Ejercicios y problemas resueltos de

Inferencia Estadística

In English

Puede encontrarse algo de teoría en [1] y [2].

Probabilidad de sucesos

Ejercicio 1ps

Supón que diriges un banco donde las cantidades de depósitos y reintegros diarios están dados por variables aleatorias independientes con distribución normal. Para los depósitos, la media es \$12.000 y la desviación estándar es \$4.000; para los reintegros, la media es \$10.000 y la desviación estándar es \$5.000.

(a) Para una semana, calcular o acotar la probabilidad de que los cinco reintegros sumen más de \$55.000
(b) Para un día particular, calcular o acotar la probabilidad de que los reintegros excedan a los depósitos en más de \$5.000

Imagina que vais a lanzar un nuevo producto mensual. El estudio prospectivo indica que los beneficios (en millones de dólares) se comportan como la cantidad aleatoria Q = (X+1)/2,325, donde X sigue una distribución t de Student con veinte grados de libertad.

(c) Para un mes particular, calcular o acotar la probabilidad de que los beneficios sean menores a uno (un millón de dólares).

(La mitad del enunciado de este ejercicio ha sido tomada del libro *Business Statistics*, Douglas Downing y Jeffrey Clark, Barron's.)

Identificación de variables y distribuciones: Del enunciado sabemos que

$$D \sim N(\mu_D = 12.000\,,\, \sigma_D^2 = 4.000^2) \quad \text{y} \quad W \sim N(\mu_W = 10.000\,,\, \sigma_W^2 = 5.000^2)$$

donde D y W representan las variables aleatorias *depósitos diarios* and *reintegros diarios*, respectivamente. (Para evitar posibles malentendidos futuros, desde el principio escribimos las varianzas –no las desviaciones estándar– en las expresiones de las distribuciones.)

(a) Como las variables son diarias, para una semana tenemos cinco medidas de ellas (una por cada día laborable).

Traducción al lenguaje matemático: Se nos pregunta por la probabilidad

$$P(W_1 + W_2 + W_3 + W_4 + W_5 > 55.000) = P\left(\sum_{i=1}^{5} W_i > 55.000\right)$$

Búsqueda de una distribución conocida: Para calcular o acotar esta probabilidad, necesitamos conocer la distribución de la suma o, alternativamente, relacionarla con una cantidad cuya distribución conozcamos. Utilizando las reglas que gobiernan las sumas y restas de variables normales,

$$\sum_{i=1}^{5} W_i \sim N(5\mu_W\,,\, 5\sigma_W^2).$$

Reescritura del suceso: Podemos reescribir fácilmente el suceso en términos de la versión estandarizada de esta distribución normal:

$$P\left(\sum_{i=1}^{5} W_i > 55.000\right) = P\left(\frac{\sum_{i=1}^{5} W_i - 5\mu_W}{\sqrt{5\sigma_W^2}} > \frac{55.000 - 5\mu_W}{\sqrt{5\sigma_W^2}}\right) = P\left(Z > \frac{55.000 - 50.000}{\sqrt{5 \cdot 5.000^2}}\right) = P(Z > 0{,}4472)$$

Consulta de la tabla: Finalmente, es suficiente consultar la tabla de la distribución normal estándar de Z. Por un lado, en la tabla nos proporcionan valores para los cuantiles 0,44 y 0,45, por lo que podríamos redondear el valor 0,4472 al más cercado, 0,45, o, más exactamente, vamos a acotar la probabilidad. Por otro lado, la tabla incluye las probabilidades de las colas inferiores, por lo que consideraremos el complementario de algunos sucesos. A partir de un dibujo de la función de densidad y los valores, es fácil deducir que

$$P(Z > 0{,}44) > P(Z > 0{,}4472) > P(Z > 0{,}45)$$

$$1 - P(Z \leq 0{,}44) > P(Z > 0{,}4472) > 1 - P(Z \leq 0{,}45)$$

$$1-0{,}6700 > P(Z > 0{,}4472) > 1-0{,}6736$$

$$0.3300 > P(Z > 0.4472) > 0.3264$$

Entonces,

$$\boxed{0{,}3264 < P\left(\sum_{i=1}^{5} W_i > 55.000\right) < 0{,}3300}$$

Nota: Es también posible relacionar la suma con la media muestral, y utilizar su distribución

$$P(\sum_{i=1}^{5} W_i > 55.000) = P(\frac{1}{5}\sum_{i=1}^{5} W_i > \frac{1}{5}55.000) = P(\bar{W} > 11.000)$$

Y

$$\bar{W} = \frac{1}{5}\sum_{i=1}^{5} W_i \sim N(\mu_W, \frac{\sigma_W^2}{5}) \quad \rightarrow \quad \frac{\bar{W}-\mu_W}{\sqrt{\frac{\sigma_W^2}{5}}} \sim N(0,1).$$

(b) Traducción al lenguaje matemático: Se nos pide la probabilidad $P(W > D + 5.000)$

Búsqueda de una distribución conocida: Para calcular o acotar la probabilidad, reescribimos el suceso para que todas las cantidades aleatorias estén en el miembro izquierdo de la desigualdad:

$$P(W > D + 5.000) = P(W - D > 5.000)$$

Ahora necesitamos conocer la distribución de $W - D$ o, alternativamente, alguna cantidad que involucra a esta diferencia y cuya distribución es conocida. Utilizando de nuevo las reglas que gobiernan las sumas y restas de variables normales:

$$W - D \sim N(\mu_W - \mu_D, \sigma_W^2 + \sigma_D^2) = N(-2.000, 5.000^2 + 4.000^2).$$

Reescritura del suceso: Podemos expresar fácilmente el suceso en términos de la versión estandarizada de esta distribución normal:

$$P(W - D > 5.000) = P\left(\frac{(W-D)-(-2.000)}{\sqrt{25\cdot 10^6 + 16\cdot 10^6}} > \frac{5.000-(-2.000)}{\sqrt{25\cdot 10^6 + 16\cdot 10^6}}\right) = P(Z > \frac{7\cdot 10^3}{\sqrt{25+16}\cdot 10^3}) = P(Z > 1{,}0932)$$

Consulta de la tabla: Finalmente,

$$P(Z > 1{,}0900) > P(Z > 1{,}0932) > P(Z > 1{,}1000)$$

$$1 - P(Z \leq 1{,}0900) > P(Z > 1{,}0932) > 1 - P(Z \leq 1{,}1000)$$

$$1 - 0{,}8621 > P(Z > 1{,}0932) > 1 - 0{,}8643$$

$$0{,}1379 > P(Z > 1{,}0932) > 0{,}1357$$

Entonces,

$$\boxed{0{,}1357 < P(W > D + 5.000) < 0{,}1379}$$

(c) Traducción al lenguaje matemático: Nos preguntan por $P\left(\frac{X+1}{2{,}325}\cdot 10^6 < 1\cdot 10^6\right) = P\left(\frac{X+1}{2{,}325} < 1\right)$

Búsqueda de una distribución conocida: No conocemos la distribución de *(X+1)/2,325*, pero sabemos que

$$X \sim t_{20}$$

Reescritura del suceso: Podemos reescribir fácilmente el suceso en términos de esa distribución conocida:

$$P\left(\frac{X+1}{2{,}325} < 1\right) = P(X + 1 < 2{,}325) = P(X < 2{,}325 - 1) = P(X < 1{,}325)$$

Consulta de la tabla: Finalmente, es suficiente consultar la tabla de la distribución *t* de Student. La cantidad 1,325 está en la tabla, y nuestra tabla proporciona las probabilidades de las colas inferiores, por lo que

$$P\left(\frac{X+1}{2{,}325}\cdot 10^6 < 1\cdot 10^6\right)=0{,}900$$

Ejercicio 2ps

Cuando un proceso de producción está funcionando correctamente, la resistencia (en ohmnios) de los componentes que produce sigue una distribución normal con desviación típica 4,68. Se toma una muestra aleatoria simple de cuatro componentes. ¿Cuál es la probabilidad de que la cuasivarianza muestral sea mayor que 30?

Variable

$R \equiv$ Resistencia (de un componente) $\qquad R \sim N(\mu, \sigma = 4{,}68)$

Muestra y estadístico

R_1, R_2, R_3, R_4 (Se mide la resistencia en cuatro componentes distintos.)

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(R_i-\bar{R})^2 \qquad \text{Cuasivarianza muestral}$$

Suceso y probabilidad

La probabilidad por la que nos preguntan es

$$P(S^2>30)$$

Para calcular la probabilidad de un suceso, tenemos que conocer la distribución de la variable aleatoria involucrada. En este caso no conocemos la distribución de S^2, aunque sabemos que como R sigue una distribución normal:

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$$

Entonces, dado que $n = 4$ y completando la desigualdad con las constantes necesarias:

$$P(S^2>30)=P\left(\frac{(n-1)S^2}{\sigma^2}>\frac{(n-1)30}{\sigma^2}\right)=P\left(X>\frac{(4-1)30}{4{,}68^2}\right)=P(X>4{,}11)$$

donde $X \sim \chi^2_3$. Por tanto, vemos que *la idea importante del ejercicio es escribir el suceso que nos piden y operar para conseguir una cantidad en la que aparezca la cuasivarianza y cuya distribución sea conocida*.

Tabla de la distribución ji-cuadrado

Como n–1=4–1=3, es suficiente mirar la tercera fila.

DF	p = .005	.01	.025	.05	.25	.5	.75	.9	.95	.975	.99
1	.000	.000	.001	.004	.10	.45	1.32	2.71	3.84	5.02	6.64
2	.010	.020	.051	.10	.58	1.39	2.77	4.61	5.99	7.38	9.21
3	.072	.11	.22	.35	1.21	2.37	4.11	6.25	7.81	9.35	11.3
4	.21	.30	.48	.71	1.92	3.36	5.39	7.78	9.49	11.1	13.3

Las probabilidades de la tabla corresponden a sucesos de la forma $P(X<\chi^2_p)$, (o $P(X\leq\chi^2_p)$, dado que la distribución es continua), así que hay que considerar el complementario:

$$P(X>4{,}11)=1-P(X\leq 4{,}11)=1-0{,}75=0{,}25$$

Ejercicio 3ps

Calcular las siguientes probabilidades:

(a) P(X = 5), donde X sigue una distribución binomial con parámetros 10 y 0,2. Busca el valor en la tabla y comprueba que es correcto utilizando la función de masa de la distribución.
(b) P(X > 2), donde X sigue una distribución de Poisson con parámetro λ = 2,7. ¿Es más fácil considerar el suceso complementario?

(a) Si la tabla que estás utilizando da probabilidades individuales $P(X = x)$, basta buscar la probabilidad que corresponde a los valores de los parámetros $k = 10$ y $p = 0{,}2$: $P(X = 5) = 0{,}0264$. Si la tabla da probabilidades acumuladas $P(X \leq x)$, debe reescribirse el suceso como $\{X=5\}=\{X\leq 5\}-\{X\leq 4\}$, por lo que

$$P(X=5)=P(X\leq 5)-P(X\leq 4)=0{,}0328-0{,}0064=0{,}0264.$$

Si no tuviésemos ninguna tabla, podríamos aplicar la definición de la función de masa:

$$P(X=5)=f(5)=\binom{10}{5}0.2^5(1-0.2)^{10-5}=252\cdot 0{,}16^5=0{,}0264.$$

(b) Si la tabla proporciona las probabilidades acumuladas de las colas inferiores $P(X \leq x)$, debe considerarse el complementario del suceso: $\{X>2\}=\{X\leq 2\}^c$, de donde

$$P(X>2)=1-P(X\leq 2)=1-0{,}4936=0{,}5064.$$

Estimación puntual

Propiedades de estadísticos y estimadores

Ejercicio 1ep-p

Para estudiar una población, consideramos un estadístico T que utiliza la información contenida en la muestra aleatoria simple $\boldsymbol{X} = (X_1, X_2,...,X_n)$*, donde el modelo poblacional X sigue una distribución ji-cuadrado con tres grados de libertad. Si*

$$T(\boldsymbol{X})=T(X_1, X_2, \ldots, X_n)=2\bar{X}-1,$$

calcular su esperanza y su varianza. Como estimador del doble de la media de la ley poblacional, ¿es T un estimador consistente en media cuadrática? Calcular el error cuadrático medio de T.

Pista: Si X sigue una distribución ji-cuadrado con m grados de libertad, $\boldsymbol{E}(X) = m$ *y* $Var(X) = 2m$.

Para calcular el valor de estas dos propiedades de la distribución en el muestreo del estadístico *T*, tenemos que aplicar las propiedades de la esperanza y de la varianza de las distribuciones de probabilidad. El conocimiento sobre la distribución de *X* se utiliza en los últimos pasos.

Esperanza:

$$\boldsymbol{E}(T(\boldsymbol{X}))=\boldsymbol{E}\left(2\left[\frac{1}{n}\sum_{i=1}^n X_i\right]-1\right)=\boldsymbol{E}\left(\frac{2}{n}\sum_{i=1}^n X_i\right)-\boldsymbol{E}(1)=\frac{2}{n}\boldsymbol{E}\left(\sum_{i=1}^n X_i\right)-1$$

$$=\frac{2}{n}\sum_{i=1}^n \boldsymbol{E}(X_i)-1=\frac{2}{n}n\,\boldsymbol{E}(X)-1=2\cdot 3-1=5$$

Varianza:

$$Var(T(\boldsymbol{X})) = Var\left(2\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - 1\right) = Var\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{2}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{4}{n^2} n\, Var(X) = \frac{4}{n} 2\cdot 3 = \frac{24}{n}$$

Independencia de $\mathbf{X_i}$ (muestra aleatoria simple)

Consistencia:

Aunque la varianza de T tiende a cero cuando n crece, la esperanza de T no tiende a $2\,\boldsymbol{E}(X)$. Entonces, T **no es un estimador consistente en media cuadrática del doble de la media de la distribución poblacional**. A partir de esta información (par de condiciones), no se puede decir nada sobre la consistencia en media de orden 1 y la consistencia en probabilidad; estos tipos de consistencia deben ser estudiados por un camino diferente.

Error cuadrático medio: Como $b(T) = \boldsymbol{E}(T) - 2\,\boldsymbol{E}(X) = (5 - 2\cdot 3)^2 = 1$,

$$MSE(T) = 1 + \frac{24}{n}.$$

Podemos ver que $MSE(T) \to 1$ cuando $n \to \infty$.

Ejercicio 2ep-p

Una muestra aleatoria simple de tamaño n *es extraída de una población normal. La media μ puede estimarse con* $\overline{X}$*. Probar que este estimador es eficiente.*

Es necesario probar que se verifica la definición de *eficacia*:

Definición

(a) La esperanza de $\overline{X}$ es μ, esto es, $\overline{X}$ es insesgado

(b) $\overline{X}$ tiene mínima varianza, lo que sucede –debido a un resultado teórico– cuando $Var(\overline{X})$ alcanza la cota mínima teórica de Cramér-Rao

$$\frac{1}{n\cdot\boldsymbol{E}\left[\left(\frac{\partial \log[f(X;\theta)]}{\partial\theta}\right)^2\right]} \quad \text{o, bajo condiciones de regularidad[1],} \quad \frac{-1}{n\cdot\boldsymbol{E}\left[\frac{\partial^2 \log[f(X;\theta)]}{\partial\theta^2}\right]}$$

donde $f(x;\theta)$ es la función de probabilidad de la ley poblacional (en este caso θ = μ), y en $f(X;\theta)$ la variable no aleatoria x se sustituye por la variable aleatoria X (en otro caso, no es posible hablar de esperanza...).

[1]. Es necesario que $\log[f(x;\theta)]$ sea dos veces diferenciable con respecto a θ. En lo concerniente a las condiciones de regularidad, la Wikipedia refiere (http://en.wikipedia.org/wiki/Fisher_information) a la ec. (2.5.16). de

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. Springer. 2nd ed. ISBN 0-387-98502-6.

(a) La esperanza de la media muestral siempre es –para cualquier población– la media poblacional. Sin embargo, podemos hacer los cálculos de nuevo:

$$\boldsymbol{E}(\overline{X}) = \boldsymbol{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\boldsymbol{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}(X_i) = \frac{1}{n} n\, \boldsymbol{E}(X) = \boldsymbol{E}(X) = \mu$$

(b) La varianza de la media muestral siempre es –para cualquier población– la varianza poblacional dividida por *n*. Sin embargo, podemos hacer de nuevo los cálculos:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\, Var(X) = \frac{\sigma^2}{n}$$

Independencia de X_i (muestra aleatoria simple)

Por otro lado, calculamos la cota mínima teórica de Cramér-Rao paso a paso:

(1) Función

$$f(X;\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

(2) Logaritmo de la función:

$$\log[f(X;\mu)] = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{(X-\mu)^2}{2\sigma^2}}\right) = -\log(\sigma\sqrt{2\pi}) - \frac{(X-\mu)^2}{2\sigma^2}$$

(3) Derivada parcial del logaritmo de la función:

$$\frac{\partial}{\partial\mu}\left(\log[f(X;\mu)]\right) = 0 - \frac{1}{2\sigma^2} 2(X-\mu)(-1) = \frac{X-\mu}{\sigma^2}$$

(4) Esperanza del cuadrado de la derivada parcial del logaritmo de la función:

$$\boldsymbol{E}\left[\left(\frac{\partial \log[f(X;\mu)]}{\partial\mu}\right)^2\right] = \boldsymbol{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4}\boldsymbol{E}\left[(X-\mu)^2\right] = \frac{1}{\sigma^4} Var(X) = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}$$

(5) Cota mínima teórica de Cramér-Rao:

$$\frac{1}{n\cdot\boldsymbol{E}\left[\left(\frac{\partial \log[f(X;\mu)]}{\partial\mu}\right)^2\right]} = \frac{1}{n\cdot\frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$$

La varianza del estimador alcanza (es igual a) la cota; entonces, el estimador tiene varianza mínima y (b) quda probada. El cumplimiento de (a) y (b) implica que **$\bar{X}$ es un estimador eficiente de μ**.

Nota: Supongamos que se puede aplicar la segunda expresión de la cota de Cramér-Rao (válida bajo ciertas condiciones de regularidad); entonces, el paso (3) sería

$$\frac{\partial^2}{\partial\mu^2}\left(\log[f(X;\mu)]\right) = \frac{\partial}{\partial\mu}\left(\frac{X-\mu}{\sigma^2}\right) = \frac{1}{\sigma^2}\cdot(-1) = -\frac{1}{\sigma^2}$$

el paso (4) sería

$$\boldsymbol{E}\left[\frac{\partial^2 \log[f(X;\mu)]}{\partial\mu^2}\right] = \boldsymbol{E}\left[-\frac{1}{\sigma^2}\right] = -\frac{1}{\sigma^2}$$

y, finalmente, el paso (5) sería

$$\frac{-1}{n\cdot\boldsymbol{E}\left[\frac{\partial^2 \log[f(X;\mu)]}{\partial\mu^2}\right]} = \frac{-1}{n\cdot\frac{-1}{\sigma^2}} = \frac{\sigma^2}{n}$$

De este modo, habríamos obtenido el mismo resultado con cálculos más fáciles, aunque el cumplimiento de las condiciones de regularidad debe verificarse antes...

Ejercicio 3ep-p

*Para estudiar la media de una población, es decir, μ = **E**(X), se considera una muestra aleatoria simple de tamaño n. No confiamos en los datos primero y último, por lo que estamos interesados en el estadístico*

$$T(\boldsymbol{X}) = T(X_1, \ldots, X_n) = \frac{1}{n-2}\sum_{i=2}^{n-1} X_i = \frac{1}{n-2}(X_2 + X_3 + \cdots + X_{n-1}) = \frac{X_2 + X_3 + \cdots + X_{n-1}}{n-2}$$

Calcular la esperanza y la varianza del estadístico. Calcular también el límite cuando n tiende a infinito.

Deben aplicarse las propiedades básicas de la media y la esperanza:

$$\boldsymbol{E}\left(T(\boldsymbol{X})\right) = \boldsymbol{E}\left(\frac{1}{n-2}(X_2 + X_3 + \cdots + X_{n-1})\right) = \frac{1}{n-2}\left(\boldsymbol{E}(X_2) + \cdots + \boldsymbol{E}(X_{n-1})\right) = \frac{1}{n-2}(n-2)\mu = \mu$$

$$Var\left(T(\boldsymbol{X})\right) = Var\left(\frac{1}{n-2}(X_2 + X_3 + \cdots + X_{n-1})\right) = \frac{1}{(n-2)^2}\sum_{i=2}^{n-1} Var(X_i) = \frac{1}{(n-2)^2}(n-2)\sigma^2 = \frac{\sigma^2}{n-2}$$

Cuando *n* crece mucho, esto es, cuando la muestra tiene cada vez más información, los límites son

$$\lim_{n\to\infty} \boldsymbol{E}\left(T(\boldsymbol{X})\right) = \lim_{n\to\infty} \mu = \mu \qquad \text{y} \qquad \lim_{n\to\infty} Var\left(T(\boldsymbol{X})\right) = \lim_{n\to\infty} \frac{\sigma^2}{n-2} = 0$$

Esto muestra que $T(\boldsymbol{X})$ tiene algunas propiedades deseables: insesgadez y varianza evanescente, par que es equivalente a la evanescencia del error cuadrático medio, que implica la consistencia (en probabilidad).

Nota: De hecho, el estimador del enunciado is la media muestral usual cuando la muestra tiene *n*–2 datos, en vez de *n*. Cuando alguno de los dos datos quitados no es confiable, tiene sentido utilizar este estimador; en otro caso, no explota la información disponible de forma óptima. Por otro lado, la media muestral puede verse afectada por valures muy grandes o muy pequeños (*atípicos*). Para hacer robusta a la media muestral, a veces se considera el estimador estudiado después de ordenar los datos (de menor a mayor o viceversa); si $X(i)$ es el *i*-ésimo dato de la muestra reordenada:

$$\tilde{T}(\boldsymbol{X}) = T(X_1, \ldots, X_n) = \frac{1}{n-2}\sum_{i=2}^{n-1} X_{(i)} = \frac{1}{n-2}(X_{(2)} + X_{(3)} + \cdots + X_{(n-1)})$$

Este nuevo estimador robusto de la media poblacional se llama *media muestral truncada*, y cualquier proporción de datos puede dejarse fuera (en vez de dos).

Ejercicio 4ep-p

Suponga que la altura de cada estudiante sigue una distribución normal con varianza 55 centímetros. Si se considera una muestra aleatoria simple de 25 estudiantes, calcular la probabilidad de que la cuasivarianza muestral sea mayor a 64,625.

La variable principal es la altura, la distribución poblacional es la normal, el tamaño muestral es 25 (menor a treinta), y se nos pregunta por la probabilidad de un evento expresado en términos de un estadístico: $P(S^2 > 64{,}625)$. Como no conocemos la distribución en el muestreo de S^2, no podemos calcular esta probabilidad directamente. En su lugar, nada más leer «cuasivarianza muestral» deberíamos pensar en el resultado teórico

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Para aplicarlo, el suceso debe reescribirse completando algunos términos. Además, cuando la tabla de la distribución ji-cuadrado da la probabilidades de las colas inferiores $P(X \le x)$, es necesario considerar el suceso complementario:

$$P(S^2 > 64{,}625) = P\left(\frac{(25-1)S^2}{55} > \frac{(25-1)64{,}625}{55}\right) = 1 - P(X \le 28{,}2) = 1 - 0{,}75 = 0{,}25.$$

Métodos

Ejercicio 1ep-m

Una población es representada por una variable aleatoria que sigue una distribución de Poisson. Dada una muestra aleatoria simple de tamaño n*, aplicar el método de máxima verosimilitud para encontrar un estimador del parámetro θ.*

(1) Función de probabilidad: Para una variable aleatoria de Poisson ($\theta > 0$),

$$f(x;\theta)=\frac{\theta^x}{x!}e^{-\theta}, \qquad x\in\{0,1,2...\}$$

(2) Función de verosimilitud:

$$L(x_{1,}x_{2,...},x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)=\prod_{i=1}^{n}\frac{\theta^{x_i}}{x_i!}e^{-\theta}=\frac{\theta^{x_1}}{x_1!}e^{-\theta}\cdot\frac{\theta^{x_2}}{x_2!}e^{-\theta}\cdots\frac{\theta^{x_n}}{x_n!}e^{-\theta}=\frac{\theta^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}e^{-n\theta}.$$

(3) Logaritmo de la función de verosimilitud:

$$\log[L(x_{1,}x_{2,...},x_n;\theta)]=\log[\theta^{\sum_{i=1}^{n}x_i}]+\log[e^{-n\theta}]-\log[\prod_{i=1}^{n}x_i!]=(\sum_{i=1}^{n}x_i)\log[\theta]-n\theta-\log[\prod_{i=1}^{n}x_i!].$$

(4) Máximo del logaritmo de la función de verosimilitud: La distribución poblacional tiene sólo un parámetro, por lo que es necesario maximizar una función de una variable. Para encontrar los «candidatos» (valores extremos locales):

$$0=\frac{\partial}{\partial\theta}\log[L(x_{1,}x_{2,...},x_n;\theta)]=(\sum_{i=1}^{n}x_i)\frac{1}{\theta}-n \quad\rightarrow\quad \theta_0=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$$

Para verificar que el candidato es un máximo (local):

$$\frac{\partial^2}{\partial\theta^2}\log[L(x_{1,}x_{2,...},x_n;\theta)]=(\sum_{i=1}^{n}x_i)\frac{-1}{\theta^2}<0$$

dado que $x\in\{0,1,2...\} \rightarrow \sum_{i=1}^{n}x_i\geq 0$. Entonces, la segunda derivada is siempre negativa: también para θ_0. Sin embargo, si sustituimos el candidato en la derivada ($\bar{x} = 0$ solo para una «muestra muy extraña»):

$$n\frac{1}{n}(\sum_{i=1}^{n}x_i)\frac{-1}{(\bar{x})^2}=n\bar{x}\frac{-1}{(\bar{x})^2}=\frac{-n}{\bar{x}}<0.$$

(5) Estimador por el método de máxima verosimilitud:

Se obtiene después de sustituir las letras minúsculas x_i (números que representan LA muestra que tenemos) por letras mayúsculas X_i (variables aleatorias que representan CUALQUIER posible muestra que podamos tener):

$$\boxed{\hat{\theta}_{ML}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}.}$$

Ejercicio 2ep-m

Una variable aleatoria poblacional tiene la función de densidad ($\theta > 0$)

$$f(x;\theta)=\begin{cases}\dfrac{2(\theta-x)}{\theta^2} & 0\le x\le\theta\\[2ex] 0 & \textit{Otherwise}\end{cases}$$

Dada una muestra aleatoria simple de tamaño n, aplicar el método de los momentos para encontrar un estimador de θ.

(1) Momentos poblacionales centrados y momentos muestrales:

La distribución poblacional tiene sólo un parámetro, por lo que sólo se igualarán los primeros momentos.

$$\alpha_1(\theta)=\boldsymbol{E}(X)=\int_{-\infty}^{+\infty} x\,f(x;\theta)\,dx=\int_0^\theta x\,\frac{2(\theta-x)}{\theta^2}\,dx=\frac{2}{\theta^2}\left(\int_0^\theta x\,\theta\,dx-\int_0^\theta x^2\,dx\right)$$

$$=\frac{2}{\theta^2}\left(\theta\frac{x^2}{2}\Big[_0^\theta-\frac{x^3}{3}\Big[_0^\theta\right)=\frac{2}{\theta^2}\left(\theta\frac{\theta^2}{2}-\frac{\theta^3}{3}\right)=\theta\,2\frac{1}{6}=\frac{1}{3}\theta$$

$$a_1(x_1,x_2,\dots,x_n)=\frac{1}{n}\sum_{i=1}^n x_i$$

(2) Sistema de ecuaciones:

$$\alpha_1(\theta)=a_1(x_1,x_2,\dots,x_n)\quad\rightarrow\quad\frac{1}{3}\theta=\frac{1}{n}\sum_{i=1}^n x_i\quad\rightarrow\quad\theta_0=\frac{3}{n}\sum_{i=1}^n x_i=3\,\bar{x}$$

(3) Estimador por el método de los momentos:

Se obtiene después de sustituir las letras minúsculas x_i (números que representan LA muestra que tenemos) por letras mayúsculas X_i (variables aleatorias que representan CUALQUIER posible muestra que podamos tener):

$$\boxed{\hat{\theta}_{MM}=\frac{3}{n}\sum_{i=1}^n X_i=3\,\bar{X}.}$$

Ejercicio 3ep-m

Dada una población, se estudia una variable aleatoria con función de densidad (*distribución exponencial*)

$$f(x;\theta)=\begin{cases}\theta e^{-\theta x} & x\ge 0\\ 0 & x<0\end{cases}$$

Para una muestra aleatoria simple de tamaño n*, aplicar tanto el método de máxima verosimilitud como el método de los momentos para encontrar un estimador del parámetro θ.*

Método de máxima verosimilitud

(1) Función de probabilidad: La función de densidad está dada en el enunciado.

(2) Función de verosimilitud:

$$L(x_{1,}x_{2,\ldots},x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)=\prod_{i=1}^{n}\left(\theta\, e^{-\theta x_i}\right)=\theta^n e^{-\theta\sum_{i=1}^{n} x_i}$$

(3) Logaritmo de la función de verosimilitud:

$$\log[L(x_{1,}x_{2,\ldots},x_n;\theta)]=\log[\theta^n]+\log[e^{-\theta\sum_{i=1}^{n} x_i}]=n\log[\theta]-\theta\sum_{i=1}^{n} x_i$$

(4) Máximo del logaritmo de la función de verosimilitud:

La distribución poblacional tiene sólo un parámetro, por lo que es necesario maximizar una función de una variable. Para encontrar los valores extremos locales ($\bar{x}$ = 0 sólo para una «muestra muy extraña»), la condición necesaria es:

$$0=\frac{\partial}{\partial\theta}\log[L(x_{1,}x_{2,\ldots},x_n;\theta)]=n\frac{1}{\theta}-\sum_{i=1}^{n} x_i \quad\rightarrow\quad \theta_0=\frac{n}{\sum_{i=1}^{n} x_i}=\frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}=\frac{1}{\bar{x}}$$

Para verificar que el candidato es un máximo (local), la condición suficiente es:

$$\frac{\partial^2}{\partial\theta^2}\log[L(x_{1,}x_{2,\ldots},x_n;\theta)]=n\frac{-1}{\theta^2}<0$$

Por tanto, la segunda derivada es siempre negativa: también para θ_0.

(5) Estimador por el método de máxima verosimilitud:

Se obtiene después de sustituir las letras minúsculas x_i (números que representan LA muestra que tenemos) por letras mayúsculas X_i (variables aleatorias que representan CUALQUIER posible muestra que podamos tener):

$$\boxed{\hat{\theta}_{ML}=\frac{n}{\sum_{i=1}^{n} X_i}=\frac{1}{\bar{X}}.}$$

<u>Método de los momentos</u>

(1) Momentos poblacionales centrados y momentos muestrales:

La distribución poblacional tiene sólo un parámetro, por lo que se igualarán sólo los primeros momentos.

$$\alpha_1(\theta)=\boldsymbol{E}(X)=\int_{-\infty}^{+\infty} x\,f(x;\theta)\,dx=\int_0^{+\infty} x\,\theta\,e^{-\theta x}\,dx=x\,\theta\frac{1}{-\theta}\left[{}_0^{+\infty}e^{-\theta x}-\int_0^{+\infty}\theta\frac{1}{-\theta}e^{-\theta x}\,dx\right.$$

$$=-x\,e^{-\theta x}\left[{}_0^{+\infty}+\frac{1}{-\theta}e^{-\theta x}\right[{}_0^{+\infty}=x\,e^{-\theta x}\left[{}_{+\infty}^{0}+\frac{1}{\theta}e^{-\theta x}\right[{}_{+\infty}^{0}=(0-0)+(\frac{1}{\theta}-0)=\frac{1}{\theta}$$

donde esta integral definida ha sido resuelta por la *regla de integración por partes*, dado que en el <u>producto</u> las funciones $x\,\theta$ y $e^{-\theta x}$ «no son del mismo tipo» (una es un polinomio y otra una exponencial):

$$\int u(x)\cdot v'(x)\,dx=u(x)\cdot v(x)-\int u'(x)\cdot v(x)\,dx$$

La función e^x crece más rápidamente x^k, para cualquier k

con

- $u=x\,\theta \quad\rightarrow\quad u'=\theta$
- $v'=e^{-\theta x} \quad\rightarrow\quad v=\int e^{-\theta x}\,dx=\frac{1}{-\theta}e^{-\theta x}$

$$a_1(x_1, x_2, \dots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(2) Sistema de ecuaciones:

$$\alpha_1(\theta) = a_1(x_1, x_2, \dots, x_n) \quad \rightarrow \quad \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \quad \theta_0 = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

(3) Estimador por el método de los momentos:

$$\boxed{\hat{\theta}_{MM} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}.}$$

Nota: En este caso, ambos métodos proporcionan el mismo estimador.

Ejercicio 4ep-m

Una variable aletoria poblacional sigue una distribución normal. Para encontrar un estimador de los parámetros $\theta = (\mu,\sigma)$ a partir de una muestra aleatoria simple de tamaño n, aplicar:

(a) el método de máxima verosimilitud *(b) el método de los momentos*

Método de máxima verosimilitud

(1) Función de probabilidad:

La función de densidad es bien conocida:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(2) Función de verosimilitud:

$$L(x_{1,}x_{2,\dots}, x_n;\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma) = \prod_{i=1}^{n}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

(3) Logaritmo de la función de verosimilitud:

$$\log[L(x_{1,}x_{2,\dots}, x_n;\mu,\sigma)] = -n\log[\sigma\sqrt{2\pi}] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

(4) Máximo del logaritmo de la función de verosimilitud:

La distribución poblacional tiene dos parámetros, por lo que is necesario maximizar una función de dos variables. Para encontrar los valores extremos locales, las condiciones necesarias son:

$$\begin{cases} \frac{\partial}{\partial\mu}\log[L(x_{1,}x_{2,\dots}, x_n;\mu,\sigma)] = 0 \\ \frac{\partial}{\partial\sigma}\log[L(x_{1,}x_{2,\dots}, x_n;\mu,\sigma)] = 0 \end{cases} \rightarrow \begin{cases} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[2(x_i-\mu)(-1)] = 0 \\ -\frac{n}{\sigma\sqrt{2\pi}}\sqrt{2\pi} - \frac{1}{2}[\sum_{i=1}^{n}(x_i-\mu)^2]\left(\frac{-2\sigma}{\sigma^4}\right) = 0 \end{cases}$$

$$\begin{cases} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \\ -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0 \end{cases} \rightarrow \begin{cases} \sum_{i=1}^{n}(x_i-\mu)=0 \\ -n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=0 \end{cases} \rightarrow \begin{cases} \sum_{i=1}^{n}x_i=n\mu \\ \sum_{i=1}^{n}(x_i-\mu)^2=n\sigma^2 \end{cases}$$

$$\begin{cases} \mu=\frac{1}{n}\sum_{i=1}^{n}x_i \\ \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \end{cases} \rightarrow \begin{cases} \mu=\bar{x} \\ \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=s_x^2 \end{cases} \rightarrow \begin{cases} \mu=\bar{x} \\ \sigma=s_x \end{cases}$$

Para verificar que el candidato es un máximo (local), las condiciones suficientes sobre las derivadas parciales de segundo orden son:

$$A=\frac{\partial^2}{\partial\mu^2}\log[L(x_{1,\dots},x_n;\mu,\sigma)]=\frac{\partial}{\partial\mu}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right]=\frac{1}{\sigma^2}\sum_{i=1}^{n}(-1)=-\frac{n}{\sigma^2}$$

$$B=\frac{\partial^2}{\partial\mu\,\partial\sigma}\log[L(x_{1,\dots},x_n;\mu,\sigma)]=\frac{\partial}{\partial\sigma}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\right]=\frac{-2\sigma}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)=-\frac{2}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu),$$

$$C=\frac{\partial^2}{\partial\sigma^2}\log[L(x_{1,\dots},x_n;\mu,\sigma)]=\frac{\partial}{\partial\sigma}\left[-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2\right]=\frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

Antes de calcular $D = B^2–AC$, el punto $(\mu,\sigma)=(\bar{x},s_x)$ se sustituye en A, B y C:

$$A|_{(\bar{x},s_x)}=-\frac{n}{s_x^2}<0$$

$$B|_{(\mu,s_x)}=-\frac{2}{s_x^3}\sum_{i=1}^{n}(x_i-\bar{x})=0 \qquad \rightarrow \qquad D|_{(\bar{x},s_x^2)}=-\left(-\frac{n}{s_x^2}\right)\left(-\frac{2n}{s_x^2}\right)=-\frac{2n^2}{s_x^4}<0$$

$$C|_{(\bar{x},s_x)}=\frac{n}{s_x^2}-\frac{3}{s_x^4}\sum_{i=1}^{n}(x_i-\mu)^2=-\frac{2n}{s_x^2}$$

dado que $\sum_{i=1}^{n}(x_i-\bar{x})=0$ y $\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=n\,s_x^2$. Entonces, $\log[L(x_1,x_{2,\dots},x_n;\mu,\sigma)]$ tiene un máximo en $(\mu,\sigma)=(\bar{x},s_x)$ porque es un valor extremos local y, en ese punto, $D < 0$ y $A < 0$.

(5) Estimador de máxima verosimilitud:

Se obtiene después de sustituir las letras minúsculas x_i (números que representan LA muestra que tenemos) por letras mayúsculas X_i (variables aleatorias que representan CUALQUIER posible muestra que podamos tener):

$$\boxed{\hat{\theta}_{ML}=\begin{cases} \hat{\mu}_{ML}=\bar{X} \\ \hat{\sigma}_{ML}=s_X \end{cases}}$$

Método de los momentos

(1) Momentos poblacionales centrados y momentos muestrales:

La población tiene dos parámetros, por lo que se igualarán los dos primeros momentos.

$$\alpha_1(\mu,\sigma)=\boldsymbol{E}(X)=\mu \qquad \alpha_2(\mu,\sigma)=\boldsymbol{E}(X^2)=Var(X)+\boldsymbol{E}(X)^2=\sigma^2+\mu^2$$

$$a_1(x_1,x_2,\dots,x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad a_2(x_1,x_2,\dots,x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i^2$$

(2) Sistema de ecuaciones:

$$\begin{cases}\alpha_1(\mu,\sigma)=a_1(x_1,x_2,\ldots,x_n)\\ \alpha_2(\mu,\sigma)=a_2(x_1,x_2,\ldots,x_n)\end{cases}\ \rightarrow\ \begin{cases}\mu=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}\\ \sigma^2+\mu^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2\end{cases}\ \rightarrow\ \begin{cases}\mu=\bar{x}\\ \sigma^2=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right)-\bar{x}^2=s_x^2\end{cases}$$

(3) Estimador por el método de los momentos:

$$\hat{\theta}_{MM}=\begin{cases}\hat{\mu}_{MM}=\bar{X}\\ \hat{\sigma}_{MM}=s_X\end{cases}$$

Nota: En este caso, ambos métodos proporcionan el mismo estimador.

Ejercicio 5ep-m

Imagina que la variable poblacional en la que estamos interesados sigue una distribución binomial, esto es, tiene una función de masa dada por

$$f(x;k,p)=\binom{k}{x}p^x(1-p)^{k-x}$$

Aplica el método de máxima verosimilitud para encontrar un estimador del parámetro p.

Pista: En los cálculos: (i) Supón que k es conocido, por lo que el parámetro de interés es θ = p; (ii) En la función de verosimilitud, agrupa los factores combinatorios en un producto y utiliza la letra A para representarlo; nótese que este producto no depende del parámetro p.

(1) Función de verosimilitud:

$$L(x_{1,}x_{2,\ldots},x_n;k,p)=\prod_{i=1}^{n}f(x_i;k,p)=\prod_{i=1}^{n}\binom{k}{x_i}p^{x_i}(1-p)^{k-x_i}=\left[\prod_{i=1}^{n}\binom{k}{x_i}\right]p^{\sum_{i=1}^{n}x_i}(1-p)^{nk-\sum_{i=1}^{n}x_i}.$$

(2) Logaritmo de la función de verosimilitud:

$$\log[L(x_{1,}x_{2,\ldots},x_n;k,p)]=\log(A)+\log(p)\sum_{i=1}^{n}x_i+\log(1-p)\left(nk-\sum_{i=1}^{n}x_i\right).$$

(3) Máximo del logaritmo de la función de verosimilitud: Para encontrar los «candidatos» (valores extremos):

$$0=\frac{\partial}{\partial p}\log[L(x_{1,}x_{2,\ldots},x_n;k,p)]=0+\frac{1}{p}\sum_{i=1}^{n}x_i-\frac{1}{1-p}\left(nk-\sum_{i=1}^{n}x_i\right)\ \rightarrow\ \frac{1-p}{p}\sum_{i=1}^{n}x_i=nk-\sum_{i=1}^{n}x_i$$

$$\frac{1}{p}-1=\frac{nk}{\sum_{i=1}^{n}x_i}-1\ \rightarrow\ p_0=\frac{1}{k}\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{k}\bar{x}$$

Para verificar que el candidato es un máximo (local):

$$\frac{\partial^2}{\partial p^2}\log[L(x_{1,}x_{2,\ldots},x_n;k,p)]=-\frac{1}{p^2}\sum_{i=1}^{n}x_i-\frac{1}{(1-p)^2}\left(nk-\sum_{i=1}^{n}x_i\right)<0$$

dado que $x\in\{0,1,2,\ldots,k\}\ \rightarrow\ nk-\sum_{i=1}^{n}x_i\geq 0$. Entonces, la segunda derivada es siempre negativa.

(4) Estimador máximo-verosímil:

$$\boxed{\hat{\theta}_{ML} = \frac{1}{k}\bar{X}\,.}$$

Ejercicio 6ep-m

La distribución uniforme U[0,θ] tiene

$$f(x;\theta) = \begin{cases} \dfrac{1}{\theta} & \text{if} \quad x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

como función de densidad. Aplicar el método de los momentos para obtener un estimador de θ.

(1) Momentos poblacionales centrados y momentos muestrales:

$$\alpha_1(\theta) = \boldsymbol{E}(X) = \int_{-\infty}^{+\infty} x\, f(x;\theta)\, dx = \int_0^{\theta} x \frac{1}{\theta} dx = \frac{1}{\theta}\frac{x^2}{2}\Big[_0^{\theta} = \frac{1}{\theta}\frac{\theta^2}{2} = \frac{\theta}{2}\,.$$

(2) Sistema de ecuaciones:

$$\alpha_1(\theta) = a_1(x_1, x_2, \ldots, x_n) \quad \rightarrow \quad \frac{\theta}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \quad \theta_0 = \frac{2}{n}\sum_{i=1}^{n} x_i = 2\,\bar{x}$$

(3) Estimador:

$$\boxed{\hat{\theta}_{MM} = \frac{2}{n}\sum_{i=1}^{n} X_i = 2\,\bar{X}\,.}$$

Intervalos de confianza

Ejercicio 1ic

Aplicar el método de la cantidad pivotal para obtener los siguientes intervalos de confianza:

- *(a) Una población normal: para μ cuando σ es conocida*
- *(b) Una población normal: para μ cuando σ es desconocida*
- *(c) Una población normal: para σ cuando μ es conocida*
- *(d) Una población normal: para σ cuando μ es desconocida*
- *(e) Dos poblaciones normales (independientes): para $\mu_x - \mu_y$ cuando σ_x y σ_y son conocidas*
- *(f) Dos poblaciones normales (independientes): para $\mu_x - \mu_y$ cuando σ_x y σ_y son desconocidas e iguales*
- *(g) Dos poblaciones normales (independientes): para σ_x/σ_y cuando μ_x y μ_y son desconocidas*
- *(h) Una población cualquiera: para μ*
- *(i) Dos poblaciones (independientes) cualesquiera: para $\mu_x - \mu_y$*
- *(j) Una población de Bernouilli: para p*
- *(k) Dos poblaciones de Bernouilli (independientes): para $p_x - p_y$*

(a) El pivote es $T(\boldsymbol{X};\mu) = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, por lo que

$$1-\alpha = P\left(a_{1-\alpha/2} < T(\boldsymbol{X};\mu) < a_{\alpha/2}\right) = P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < +z_{\alpha/2}\right) = P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < +z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = P\left(\bar{X}+z_{+\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

(b) El pivote es $T(\boldsymbol{X};\mu) = \dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$, por lo que

$$1-\alpha = P\left(1-a_{\alpha/2} < T(\boldsymbol{X};\mu) < a_{\alpha/2}\right) = P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < +t_{\alpha/2}\right) = P\left(-t_{\alpha/2}\frac{S}{\sqrt{n}} < \bar{X}-\mu < +t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$= P\left(-\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < -\mu < -\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = P\left(\bar{X}+t_{+\alpha/2}\frac{S}{\sqrt{n}} > \mu > \bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$= P\left(\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

(c) El pivote es $T(\boldsymbol{X};\sigma) = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\dfrac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$, por lo que

$$1-\alpha = P\left(a_{1-\alpha/2} < T(\boldsymbol{X};\sigma) < a_{\alpha/2}\right) = P\left(\chi_{1-\alpha/2} < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} < \chi_{\alpha/2}\right)$$

$$= P\left(\frac{\chi_{1-\alpha/2}}{\sum_{i=1}^{n}(X_i-\mu)^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}}{\sum_{i=1}^{n}(X_i-\mu)^2}\right) = P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{1-\alpha/2}}\right)$$

(d) El pivote es $T(\boldsymbol{X};\sigma) = \dfrac{ns^2}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, por lo que

$$1-\alpha = P\left(a_{1-\alpha/2} < T(\boldsymbol{X};\sigma) < a_{\alpha/2}\right) = P\left(\chi_{1-\alpha/2} < \frac{ns^2}{\sigma^2} < \chi_{\alpha/2}\right) = P\left(\frac{\chi_{1-\alpha/2}}{ns^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}}{ns^2}\right)$$

$$= P\left(\frac{ns^2}{\chi_{1-\alpha/2}} > \sigma^2 > \frac{ns^2}{\chi_{\alpha/2}}\right) = P\left(\frac{ns^2}{\chi_{\alpha/2}} < \sigma^2 < \frac{ns^2}{\chi_{1-\alpha/2}}\right)$$

(e) El pivote es $T(\boldsymbol{X},\boldsymbol{Y};\mu_X\mu_Y) = \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$, por lo que

$$1-\alpha = P\left(a_{1-\alpha/2} < T(\boldsymbol{X},\boldsymbol{Y};\mu_X\mu_Y) < a_{\alpha/2}\right) = P\left(-z_{\alpha/2} < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} < +z_{\alpha/2}\right)$$

$$= P\left(-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} < (\bar{X}-\bar{Y})-(\mu_X-\mu_Y) < +z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}\right)$$

$$= P\left(-(\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} < -(\mu_X-\mu_Y) < -(\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} \right)$$

$$= P\left((\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} > (\mu_X-\mu_Y) > (\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} \right)$$

$$= P\left((\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} < (\mu_X-\mu_Y) < (\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}} \right)$$

(f) El pivote es $T(\boldsymbol{X},\boldsymbol{Y};\mu_X,\mu_Y) = \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_p^2\left(\dfrac{1}{n_X}+\dfrac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2}$, donde se involucra la varianza muestral ponderada $S_p^2 = \dfrac{n_X s_X^2 + n_Y s_Y^2}{n_X+n_y-2} = \dfrac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_y-2}$, por lo que

$$1-\alpha = P(a_{1-\alpha/2} < T(\boldsymbol{X},\boldsymbol{Y};\mu_X,\mu_Y) < a_{\alpha/2}) = P\left(-t_{\alpha/2} < \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_p^2\left(\dfrac{1}{n_X}+\dfrac{1}{n_Y}\right)}} < +t_{\alpha/2} \right)$$

$$= P\left(-t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)} < (\bar{X}-\bar{Y})-(\mu_X-\mu_Y) < +t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)} \right)$$

$$= \cdots = P\left((\bar{X}-\bar{Y}) - t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)} < (\mu_X-\mu_Y) < (\bar{X}-\bar{Y}) + t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)} \right)$$

(g) El pivote es $T(\boldsymbol{X},\boldsymbol{Y};\sigma_X,\sigma_Y) = \dfrac{\dfrac{S_X^2}{\sigma_X^2}}{\dfrac{S_Y^2}{\sigma_Y^2}} = \dfrac{S_X^2\sigma_Y^2}{\sigma_X^2 S_Y^2} \sim F_{n_X-1,\,n_Y-1}$, por lo que

$$1-\alpha = P(a_{1-\alpha/2} < T(\boldsymbol{X};\sigma) < a_{\alpha/2}) = P\left(f_{1-\alpha/2} < \frac{S_X^2\sigma_Y^2}{\sigma_X^2 S_Y^2} < f_{\alpha/2} \right) = P\left(f_{1-\alpha/2}\frac{S_Y^2}{S_X^2} < \frac{\sigma_Y^2}{\sigma_X^2} < f_{\alpha/2}\frac{S_Y^2}{S_X^2} \right)$$

$$= \cdots = P\left(\frac{1}{f_{\alpha/2}}\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{1}{f_{1-\alpha/2}}\frac{S_X^2}{S_Y^2} \right)$$

(h) El pivote es $T(\boldsymbol{X};\mu) = \dfrac{\bar{X}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}} \rightarrow N(0,1)$, por lo que cuando $n > 30$ se puede aplicar el intervalo obtenido en (a) a cualquier población. Además, si σ^2 es desconocida y $n > 100$ cualquiera de s^2 ó S^2 puede utilizarse en su lugar.

(i) El pivote es $T(\boldsymbol{X},\boldsymbol{Y};\mu_X,\mu_Y) = \dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X}+\dfrac{\sigma_Y^2}{n_Y}}} \rightarrow N(0,1)$, por lo que cuando $n_X > 30$ y $n_Y > 30$ se puede aplicar el intervalo obtenido en (e) a dos poblaciones (independientes) cualesquiera. Además, si σ_X^2 y

σ_Y^2 son desconocidas y $n_X > 100$ y $n_Y > 100$, sus estimadores muestrales pueden utilizarse en su lugar.

(j) El pivote es $T(\boldsymbol{X}; p) = \dfrac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \rightarrow N(0,1)$, por lo que

$$1-\alpha = P(a_{1-\alpha/2} < T(\boldsymbol{X}; p) < a_{\alpha/2}) \approx P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} < +z_{\alpha/2} \right)$$

$$= \cdots = P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

(k) El pivote es $T(\boldsymbol{X}, \boldsymbol{Y}; p_X, p_Y) = \dfrac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}} \rightarrow N(0,1)$, por lo que

$$1-\alpha = P(a_{1-\alpha/2} < T(\boldsymbol{X}, \boldsymbol{Y}; p_X, p_Y) < a_{\alpha/2}) \approx P\left(-z_{\alpha/2} < \frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \dfrac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}} < +z_{\alpha/2} \right) = \cdots =$$

$$= P\left((\hat{p}_X - \hat{p}_Y) - z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} < (p_X - p_Y) < (\hat{p}_X - \hat{p}_Y) + z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}} \right)$$

Poblaciones normales

Ejercicio 1ic-e

Para estimar la altura media de los árboles de un bosque, se considera una muestra aleatoria simple con 20 elementos, proporcionando

$$\bar{x} = 14{,}70\,u \quad \text{and} \quad s = 6{,}34\,u\,,$$

donde u denota una unidad de longitud y s^2 es la cuasivarianza muestral. Si se supone que la variable poblacional altura *es normal, encuentra un intervalo de confianza del 95 por ciento. ¿Cuál es el margen de error?*

A partir de la información del enunciado, sabemos que la variable se distribuye normalmente y tiene varianza desconocida. El tamaño muestral es n = 20 (menor a 30, por lo que ningún resultado asintótico podría ser utilizado). Para aplicar el método de la cantidad pivotal, necesitamos un pivote con distribución conocida, fácil de manipular y con μ involucrado en su expresión. Consultando una tabla de estadísticos (p.ej. en [2]),

$$T(\boldsymbol{X}, \mu) = \frac{\bar{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1}$$

donde $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple, S^2 es la cuasivarianza muestral y t_k denota la distribución t de Student con k grados de libertad (en otras expresiones t_p es el cuantil de probabilidad p), el intervalo se construye como sigue

$$1-\alpha = P\left(a_{1-\alpha/2} < T(\boldsymbol{X};\mu) < a_{\alpha/2}\right) = P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < +t_{\alpha/2}\right) = P\left(-t_{\alpha/2}\frac{S}{\sqrt{n}} < \bar{X}-\mu < +t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$= P\left(-\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < -\mu < -\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = P\left(\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}} > \mu > \bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

$$= P\left(\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right) \quad \rightarrow \quad I = \left[\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right]_{1-\alpha} = \boxed{\bar{X}\mp t_{\alpha/2}\frac{S}{\sqrt{n}}}$$

Nota: Las cantidades $t_{\alpha/2}$ y S también dependen del tamaño muestral n.

Para utilizar esta fórmula general con los datos específicos que tenemos, se necesitan los cuantiles de la distribución con $k = n–1 = 20–1 = 19$ grados de libertad

$$95\% \ \rightarrow \ 0{,}95 = 1–\alpha \ \rightarrow \ \alpha = 0{,}05$$

En la tabla de la distribución t, debemos buscar el cuantil dado para $p = 1–\alpha+\alpha/2 = 1–\alpha/2 = 0{,}975$ para una tabla de probabilidades de colas inferiores, o $p = \alpha/2 = 0{,}025$ para una tabla de probabilidades de cola superior; si se utiliza una tabla de dos colas, debe considerarse el cuantil dado para $p = 1–\alpha = 0{,}950$. Cualquiera que sea la tabla utilizada, el cuantil es 2,093. Finalmente,

$$I_0 = \bar{x}\mp t_{0{,}05/2}\frac{s}{\sqrt{20}} = 14{,}70\,u \mp 2{,}093\,6{,}34\,\frac{u}{\sqrt{20}} = 14{,}70\,u \mp 2{,}97\,u$$

Aplicando la definición del margen de error a este intervalo,

$$ME = t_{\alpha/2}\frac{S}{\sqrt{n}} = 2{,}093\,6{,}34\,\frac{u}{\sqrt{20}} = 2{,}97\,u$$

Ejercicio 2ic-e

El número de unidades demandadas *de un producto se modeliza, para dos áreas diferentes e independientes A y B, por las distribuciones* $N(\mu_A,\sigma^2)$ *y* $N(\mu_B,\sigma^2)$*, respectivamente. Para estudiar la diferencia entre las medias, se consideran las siguientes muestras aleatorias simples*

Area A	*4*	*7*	*7*	*4*	*8*
Area B	*7*	*6*	*7*	*7*	*8*

Encontrar un intervalo de confianza 99 por ciento. ¿Cuál es el margen de error?

Elección del pivote adecuado: Hay dos poblaciones normales independientes, estamos interesados en $\mu_A–\mu_B$ y las varianzas son desconocidas pero iguales. Entonces, a partir de una tabla de estadísticos (p.ej. en [2]), se selecciona el pivote

$$T(\boldsymbol{X},\boldsymbol{Y};\mu_X,\mu_Y) = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2},$$

de la talba, con *varianza muestral ponderada* $S_p^2 = \frac{n_X s_X^2 + n_Y s_Y^2}{n_X+n_y-2} = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_y-2}$.

Método de la cantidad pivotal: Véase Ejercicio 1ic.

Cálculos:

- $\bar{x}=\frac{1}{5}\sum_{i=1}^{5} x_i=\frac{4+7+7+4+8}{5}=6$ y $\bar{y}=\frac{1}{5}\sum_{i=1}^{5} x_i=\frac{7+6+7+7+8}{5}=7$
- $S_x^2=\frac{1}{5-1}\sum_{i=1}^{5}(x_i-6)^2=\frac{(4-6)^2+\cdots+(8-6)^2}{4}=3{,}5$ y $S_y^2=\frac{1}{5-1}\sum_{i=1}^{5}(y_i-7)^2=0{,}5$

  por lo que $S_p^2=\frac{(5-1)3.5+(5-1)0.5}{5+5-2}=2.$

  Si la tabla proporciona probabilidades de colas inferiores, es necesario buscar $t_{8,1-0,005}$

- Dado que $1-\alpha=0{,}99$, $t_{n_x+n_y-2,\frac{\alpha}{2}}=t_{5+5-2,\frac{0{,}01}{2}}=t_{8,0{,}005}=3{,}355.$

Intervalo de confianza:

$$CI_\alpha=\left[(6-7)-3{,}355\sqrt{2\left(\frac{1}{5}+\frac{1}{5}\right)},(6-7)+3{,}355\sqrt{2\left(\frac{1}{5}+\frac{1}{5}\right)}\right]=-1\mp 3{,}001$$

Margen de error:

$$E=t_{n_x+n_y-2,\frac{\alpha}{2}}\sqrt{S_p^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}=3{,}355\sqrt{2\left(\frac{1}{5}+\frac{1}{5}\right)}=3{,}001$$

Ejercicio 3ic-e

La nota de una prueba de aptitud sigue una distribución normal con desviación típica 28,2. Una muestra aleatoria simple de nueve alumnos proporciona los resultados siguientes:

$$\sum_{i=1}^{9} x_i=1.098 \qquad \sum_{i=1}^{9} x_i^2=138.148$$

a) Hallar un intervalo de confianza al 90% para la media poblacional μ.
b) Razonar sin hacer cálculos si la longitud de un intervalo al 95% será menor, mayor o igual que la del obtenido en el apartado anterior.
c) ¿Cuál debería ser el tamaño de muestra mínimo necesario para obtener un intervalo del 90% de nivel de confianza con longitud (entre extremos) igual a 10?

Identificar la variable

X ≡ Nota (de un alumno) $\qquad X \sim N(\mu, \sigma^2 = 28{,}2^2)$

Información muestral

Muestra teórica (aleatoria): $X_1,\ldots, X_9$ m.a.s. (se van a tomar las notas de nueve alumnos)
Muestra numérica: $x_1,\ldots, x_9 \rightarrow \sum_{i=1}^{9} x_i=1.098 \quad \sum_{i=1}^{9} x_i^2=138.148$ (se han tomado las notas)

Podemos observar que no se conocen los valores x_i de la muestra. Sin embargo, se conoce información construida a partir de estos valores. Debe ser *suficiente* para hacer los cálculos, que deben involucrar a las sumas anteriores.

(a) Intervalo de confianza

Para elegir la fórmula adecuada con que calcular el intervalo de confianza, tenemos en cuenta que:

- El tamaño muestral, $n = 9$, es pequeño, por lo que no se podrían utilizar las fórmulas asintóticas
- Se sabe que la variable sigue una distribución normal
- Finalmente, como nos informan del valor de la desviación típica poblacional, no es necesario estimarla

A partir de una tabla de estadísticos (p.ej. En [2]), se selecciona el estadístico apropiado y, después de aplicar el método de la cantidad pivotal (véase Ejercicio 1ic), concluimos que debemos usar la expresión

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} < \mu < \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right) = 1-\alpha = 0{,}90$$

donde $z_{\alpha/2}$ es el valor de la distribución normal estándar que verifica $P(Z > z_{\alpha/2}) = \alpha/2$, es decir, el valor tal que deja un área igual a $\alpha/2$ a la derecha (cola superior). Vamos a calcular las cantidades que aparecen en la fórmula:

➔ $\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9}1.098 = 122$

➔ Un nivel de confianza del 90% implica que $\alpha = (100–90)/100 = 0{,}1$. El cuantil $z_{\alpha/2} = z_{0{,}05}$ se busca en la tabla. Como la tabla nos informa de las probabilidades de la forma $P(Z \le z_p)$, buscamos el valor $z_p = z_{1-\alpha/2} = z_{1-0{,}05} = z_{0{,}95} = -1{,}645$ Entonces $z_{\alpha/2} = 1{,}645$.

➔ Por el enunciado, $\sigma = 28{,}2$

➔ Por último, $n = 9$

El intervalo es

$$\boxed{CI_{0{,}1} = \left[122 - 1{,}65\frac{28{,}2}{\sqrt{9}},\ 122 + 1{,}65\frac{28{,}2}{\sqrt{9}} \right]}$$

(b) Longitud del intervalo

Para responder a la pregunta se puede razonar que, fijos todos los parámetros excepto la longitud, si se quiere mayor certeza es necesario ampliar el intervalo, es decir, aumentar su longitud. La manera formal de justificar esto consiste en usar la fórmula del intervalo,

$$L = \left(\bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right) - \left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right) = 2\, z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}.$$

Ahora, si σ y n permanecen fijos, para estudiar cómo varía L con α basta ver cómo varía el cuantil. Para el intervalo del 95%:

- $\alpha = (100–95)/100 = 0{,}05$ → α disminuye
- Ahora la cantidad $z_{\alpha/2}$ debe dejar menos área (probabilidad) a la derecha → $z_{\alpha/2}$ aumenta

Por tanto, de la expresión anterior se deduce que L **aumenta**.

(c) Tamaño muestral

Ahora se vuelve al intervalo del apartado primero, del 90%, y se pregunta por el valor de n para un α y una L dadas. De la expresión de la longitud es necesario despejar n

$$L = 2\, z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad \sqrt{n} = 2\, z_{\alpha/2}\frac{\sigma}{L} \quad \rightarrow \quad n = \left(2\, z_{\alpha/2}\frac{\sigma}{\sqrt{L}} \right)^2 = \left(2 \cdot 1{,}645\frac{28{,}2}{\sqrt{10}} \right)^2 = 860{,}78 \quad \rightarrow \quad \boxed{n = 861.}$$

Si se tomase un tamaño mayor que 861, se obtendría un intervalo de precisión mayor; sin embargo, en la práctica esto también implicaría un mayor gasto económico.

Ejercicio 4ic-e

Para prever la inflación en el año, se ha recogido una muestra aleatoria simple

$$1{,}5 \quad 2{,}1 \quad 1{,}9 \quad 2{,}3 \quad 2{,}5 \quad 3{,}2 \quad 3{,}0$$

Si se supone que la variable inflación sigue una distribución normal:

(a) Utilizando estos datos, construye un intervalo de confianza al 99% para la media de la inflación.
(b) Construye un intervalo de confianza al 90% para la desviación típica.
(c) Los expertos opinan que el intervalo de confianza calculado para la media es demasiado amplio, y desean una longitud total igual a 1,2 puntos. Hallar el nivel de confianza para este nuevo intervalo.

Identificar la variable

$X \equiv$ Previsión de la inflación (de un país) $\qquad X \sim ?$

Información muestral

Muestra teórica: $X_1,\ldots, X_7$ m.a.s.

Muestra númerica: $x_1,\ldots, x_7 \rightarrow 1{,}5 \quad 2{,}1 \quad 1{,}9 \quad 2{,}3 \quad 2{,}5 \quad 3{,}2 \quad 3{,}0$

En este ejercicio sí conocemos los valores x_i de la muestra.

(a) Intervalo de confianza para la media

Para elegir la fórmula adecuada del intervalo, tenemos en cuenta que:

- El tamaño muestral, $n = 7$, es pequeño, por lo que no debemos pensar en ninguna fórmula asintótica
- La desviación típica poblacional es desconocida, por lo que debe ser estimada por la cuasivarianza muestral

A partir de una tabla de estadísticos (p.ej. En [2]), se selecciona el estadístico apropiado y, después de aplicar el método de la cantidad pivotal (véase Ejercicio 1ic), concluimos que debemos usar la expresión

$$P\left(\bar{X} - t_{\alpha/2}\sqrt{\frac{S^2}{n}} < \mu < \bar{X} + t_{\alpha/2}\sqrt{\frac{S^2}{n}} \right) = 1-\alpha = 0{,}99$$

donde $t_{\alpha/2}$ es el cuantil tal que $P(T > t_{\alpha/2}) = \alpha/2$. Vamos a calcular las cantidades en la fórmula:

➔ $\bar{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = 2{,}35$

➔ El nivel de confianza es 99%, por lo que $\alpha = (100–99)/100 = 0{,}01$. El cuantil $t_{\alpha/2} = t_{0{,}01/2} = t_{0{,}005}$ se encuentra en la tabla de la distribución $t_{7\text{-}1}$ de Student. Dado que $t_p = t_{1-\alpha/2} = t_{1-0{,}005} = t_{0{,}995} = -3{,}71$, entonces $t_{\alpha/2} = 3{,}71$

➔ Utilizando la muestra, $S^2 = \frac{1}{7-1}\sum_{i=1}^{7}(x_i - \bar{x})^2 = 0{,}6^2 = 0{,}36$

➔ Por último, $n = 7$

El intervalo es

$$CI_{0{,}01} = \left[2{,}35 - 3{,}71\sqrt{\frac{0{,}60}{7}},\ 2{,}35 + 3{,}71\sqrt{\frac{0{,}60}{7}} \right]$$

(b) Intervalo de confianza para la desviación típica

A partir de una tabla de estadísticos (p.ej. En [2]), se selecciona el estadístico apropiado y, después de aplicar

el método de la cantidad pivotal (véase Ejercicio 1ic) y tomando la raíz cuadrada, concluimos que debemos usar la expresión

$$P\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{p_b}}}\leq\sigma\leq\sqrt{\frac{(n-1)S^2}{\chi^2_{p_a}}}\right)=1-\alpha=0{,}90$$

donde $\chi^2_{p_a}$ y $\chi^2_{p_b}$ son valores de la distribución ji-cuadrado, con parámetro n–1 = 7–1 = 6, tales que

$$\chi^2_{p_a} \text{ tal que } P(X<\chi^2_{p_a})=\alpha/2 \qquad \chi^2_{p_b} \text{ tal que } P(X>\chi^2_{p_b})=\alpha/2$$

Las cantidades que aparecen en la fórmula son:

- ➔ Tamaño muestral $n = 7$
- ➔ $S^2=0{,}36$
- ➔ Como $\alpha = 0{,}1$, $\chi^2_{0{,}05}=1{,}64$ and $\chi^2_{0{,}95}=12{,}6$

El intervalo es

$$CI_{0,1}=\left[\sqrt{\frac{6\cdot 0{,}36}{12{,}6}},\sqrt{\frac{6\cdot 0{,}36}{1{,}64}}\right]$$

(c) Nivel de confianza

La longitud del intervalo es

$$L=\left(\bar{X}+t_{\alpha/2}\sqrt{\frac{S^2}{n}}\right)-\left(\bar{X}-t_{\alpha/2}\sqrt{\frac{S^2}{n}}\right)=2\,t_{\alpha/2}\sqrt{\frac{S^2}{n}}$$

donde L está dada y debe encontrarse α. Sin embargo, previamente es necesario encontrar $t_{\alpha/2}$.

$$t_{\alpha/2}=\frac{L\sqrt{n}}{2S}=\frac{1\cdot\sqrt{7}}{2\cdot 0{,}6}=2{,}2.$$

A partir de la tabla, α/2 = 0,10 por lo que α = 0,20 y el nivel de confianza en tanto por ciento es:

$$100 - 0{,}20 \cdot 100 = 100 - 20 = \mathbf{80\%}.$$

Cualesquier poblaciones

Ejercicio 1ic-a

La duración media de préstamos en la biblioteca de una universidad en el curso pasado fue de veinte días. Se toma una muestra aleatoria simple de cien libros este año, y se obtienen unos valores de dieciocho y ocho días para la media y varianza muestrales, respectivamente. Construir un intervalo de confianza del 99% para la duración media de préstamos en el curso pasado.

Identificamos la variable

$X \equiv$ Duración (de un préstamo) $\qquad X \sim ?$

Información muestral

Muestra teórica: $X_1,\ldots,X_{100}$ m.a.s.

Muestra numérica: $x_1,\ldots,x_{100} \rightarrow \bar{x}=18, \quad s^2=8$

Los valores x_i de la muestra son desconocidos. A cambio, se conoce la evaluación de algunos estadísticos. Ésta debe ser *suficiente* para hacer los cálculos, y las fórmulas deben escribirse en términos de $\bar{X}$ y S^2.

Intervalo de confianza

Para elegir la fórmula adecuada con que calcular el intervalo de confianza, tenemos en cuenta que:

- El tamaño muestral es grande (>30), n = 100, por lo que se puede utilizar alguna fórmula asintótica
- La varianza poblacional es desconocida, pero se estima por la varianza muestral

A partir de una tabla de estadísticos (p.ej. En [2]), se selecciona el estadístico apropiado y, aplicando el método de la cantidad pivotal (véase Ejercicio 1ic) y sustituyendo σ por S, concluimos que debemos usar la expresión

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{S^2}{n}} < \mu < \bar{X} + z_{\alpha/2}\sqrt{\frac{S^2}{n}} \right) = 1-\alpha = 0{,}99$$

donde $z_{\alpha/2}$ es el cuantil tal que $P(Z > z_{\alpha/2}) = \alpha/2$. Calculamos las cantidades:

- ➔ Media muestral $\bar{x}=18$
- ➔ Para la confianza del 99%, $\alpha = 0.01$ y $z_{\alpha/2}=2{,}575$.
- ➔ Para calcular S^2 se utiliza la expresión $(n-1)S^2=\sum_{i=1}^{100}(x_i-\bar{x})^2=n\,s^2$,

$$S^2=\frac{n}{n-1}s^2=\frac{100}{99}8=8{,}1$$ (nótese que el factor $n/(n–1)$ tiende a 1 cuando n crece).

- ➔ Finalmente, n = 100

El intervalo es

$$CI_{0{,}01}=\left[18-2{,}575\frac{8{,}1}{\sqrt{100}},\ 18+2{,}575\frac{8{,}1}{\sqrt{100}} \right]$$

Tamaño muestral mínimo

Ejercicio 1ic-t

Para estimar la media de una distribución normal con desviación típica 5, ¿cuán grande debe ser la muestra para construir un intervalo del 95 por ciento de confianza con un margen de error igual a 1,2?

Es necesario relacionar el tamaño muestral con el margen de error y el nivel de significancia. La anchura del intervalo es el doble del margen de error, y depende de n y α. Entonces, dado que para una población normal con varianza conocida el intervalo se obtiene como en (a) de Ejercicio 1ic,

$$CI_\alpha=\left[\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right] \rightarrow E=z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n=\left(z_{\alpha/2}\frac{\sigma}{E} \right)^2=\left(1{,}96\frac{5}{1{,}2} \right)^2=66{,}7 \rightarrow \boxed{n=67}$$

porque $1-\alpha=0{,}95 \rightarrow \alpha=1-0{,}95=0{,}05 \rightarrow z_{\alpha/2}=z_{0{,}025}=1{,}96$.

Véanse también:

Ejercicio 3ic-e
Ejercicio 4ic-e

Contrastes de hipótesis

Ejercicio 1ch

Para contrastar si una moneda es justa o no, ha sido lanzada 100.000 veces, y 50.347 de ellas han sido caras. ¿Debería rechazarse, como hipótesis nula, la justicia de la moneda cuando α = 0,1?

(a) Aplicar un contraste paramétrico de significancia.
(b) Aplicar el contraste no paramétrico de bondad de ajuste de ji-cuadrado.
(c) Aplicar el contraste no paramétrico de posición de los signos.

(a) Contraste paramétrico de significancia

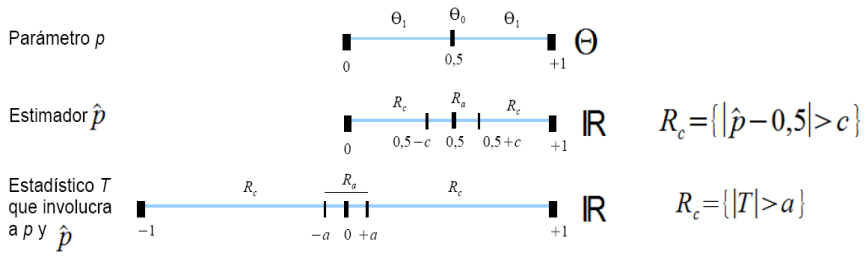
Hipótesis y nivel de significancia: Dado que debe aplicarse un contraste paramétrico, la moneda es modelizada con una variable aleatoria de Bernouilli, y las hipótesis son

$$H_0:\ p=\frac{1}{2} \quad \text{y} \quad H_1:\ p\neq\frac{1}{2}$$

Nótese que la pregunta se basa en el valor del parámetro p, mientras que se supone la distribución de Bernouilli para ambas hipótesis; en algunos contrastes no paramétricos, esta distribución no se supone incluso. Por otro lado, el nivel $\alpha = 0{,}1$ está dado.

Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]), dado que la variable poblacional es de Bernouilli y el marco asintótico puede aplicarse (porque n es grande)

$$T(\boldsymbol{X};p)=\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{d} N(0,1)$$



Como la distribución normal es simétrica, un cuantil determina los dos valores críticos. Para calcular este cuantil, se aplica la definición de error de tipo I:

$$\alpha=P(\text{Error tipo I})=P(\text{Rechazar } H_0 \mid H_0 \text{ cierta})=P(T_\theta(\boldsymbol{X})\in R_c \mid \theta\in\Theta_0)=P(|T_p(\boldsymbol{X})|>a \mid p=\frac{1}{2})$$

$$=1-P(|T_p(\boldsymbol{X})|\leq a \mid p=\frac{1}{2})\approx 1-P(|Z|\leq a) \ \rightarrow\ P(|Z|\leq a)\approx 1-\alpha=1-0{,}1=0{,}9 \ \rightarrow\ a\approx 1{,}645$$

Tabla de una o dos colas

Regla de decisión: Si aplicamos la metodología basada en la región crítica,

$$T(\boldsymbol{x};\frac{1}{2})=\frac{\frac{50.347}{100.000}-\frac{1}{2}}{\sqrt{\frac{\frac{50.347}{100.000}\left(1-\frac{50.347}{100.000}\right)}{100.000}}}=2{,}19>1{,}645 \ \rightarrow\ T(\boldsymbol{x};\frac{1}{2})\in R_c \ \rightarrow \textbf{Se rechaza } H_0$$

Si aplicamos la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta}) = P(|T_p(\boldsymbol{X})| \geq |T_p(\boldsymbol{x})| \mid p = \frac{1}{2})$$

$$\approx P(|Z| \geq 2{,}19) = 2\,P(Z \leq -2{,}19) = 2 \cdot 0{,}0143 = 0{,}0248 \;\rightarrow\; pV < 0{,}1 = \alpha \;\rightarrow \textbf{Se rechaza } H_0$$

Tabla de una cola, dado que 2,19 no está en la tabla de dos colas que tenemos

(b) Contraste no paramétrico de bondad de ajuste ji-cuadrado

Hipótesis y nivel de significancia: El nivel es $\alpha = 0{,}1$. Para un contraste no paramétrico de bondad de ajuste, la hipótesis nula supone que la muestra fue generada por una distribución de Bernouilli con $p = 1/2$, mientras que la hipótesis alternativa supone que fue generada por una distribución diferente (de Bernouilli o no).

$$H_0\colon\ X \sim Bern\left(\frac{1}{2}\right) \quad \text{y} \quad H_1\colon\ X \sim F \neq Bern\left(\frac{1}{2}\right)$$

Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]),

$$T(\boldsymbol{X}) = \sum_{i=1}^{K} \frac{(N_i - \hat{e}_i)^2}{\hat{e}_i} \overset{d}{\rightarrow} \chi^2_{K-s-1}$$

con

K = 2 clases y celdas

No ha tenido que estimarse ningún parámetro, por lo que $s = 0$

Parámetro θ: Θ_0 (0) — Θ_1 — Θ

Estadístico T: R_a — R_c — $\mathbb{R}$ (0)

$$R_c = \{T > a\}$$

Para calcular el cuantil a, se aplica la definición de error de tipo I:

$$\alpha = P(\textit{Error tipo I}) = P(\textit{Reject } H_0 \mid H_0 \textit{ true}) = P(T(\boldsymbol{X}) \in R_c \mid T \approx \chi^2_1) \approx P(\chi^2_1 > a) = 1 - P(\chi^2_1 \leq a)$$

$$\rightarrow \quad P(\chi^2_1 \leq a) \approx 1 - \alpha = 1 - 0{,}1 = 0{,}9 \quad \rightarrow \quad a \approx 2{,}71$$

Regla de decisión: Dado que $e_i = n\,p_i = n\,P_\theta(i^{th}\,class) = 100.000\,\frac{1}{2} = 50.000$

Tabla esperada (bajo H_0)		
Clases	Cara	Cruz
e_i	50000	50000

Tabla empírica (muestra)		
Clases	Cara	Cruz
n_i	50347	49653

Si aplicamos la metodología basada en la región crítica,

$$T(\boldsymbol{x}) = \sum_{i=1}^{2} \frac{(n_i - e_i)^2}{e_i} = \frac{(50.347 - 50.000)^2}{50.000} + \frac{(49.653 - 50.000)^2}{50.000} = 4{,}82 \;\rightarrow\; T(\boldsymbol{x}) \in R_c \;\rightarrow \textbf{Se rechaza } H_0$$

Si aplicamos la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ true}) = P(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid T \approx \chi^2_1) \approx P(\chi^2_1 \geq 4{,}82)$$

$$= 1 - P(\chi^2_1 < 4{,}82) < 1 - P(\chi^2_1 < 3{,}84) = 1 - 0{,}95 = 0{,}05 \;\rightarrow\; pV < 0{,}1 = \alpha \;\rightarrow \textbf{Se rechaza } H_0$$

4,82 no está en la tabla que tenemos, mientras que 3,84 está

Nota: En algunas situaciones, acotar el nivel crítico es suficiente para ver si es menor o mayor que α (para encontrar la acotación, se utiliza el valor adecuado más cercano incluido en la tabla: el menor o el mayor). Algunas veces esto no es suficiente y hay que calcular el valor exacto de la probabilidad teóricamente o utilizando algún programa.

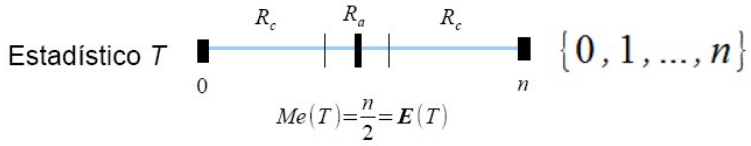
(c) Contraste no paramétrico de posición de los signos

Hipótesis y nivel de significancia: El nivel es $\alpha = 0{,}1$. Para un contraste no paramétrico, si *cara* y *cruz* se escriben equivalentemente como +1 y -1, respectivamente, las hipótesis son

$$H_0\colon\ Me(X)=0 \quad \text{y} \quad H_1\colon\ Me(X)\neq 0$$

Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]),

$$T(\boldsymbol{X}) = Número\{X_i - Me(X) > 0\}$$
$$= Número\{X_i > 0\} \sim Bin\left(n, \frac{1}{2}\right)$$



$$R_c = \{|T - n/2| > a\}$$

Para calcular el cuantil a, se aplica la definición de error de tipo I. Sin embargo, conocemos la distribución de T, pero R_c se ha escrito fácilmente en términos de $T - n/2$, cuya distribución está involucrada en un resultado asintótico bien conocido (además, las probabilidades de la binomial no están tabuladas para $n = 100.000$)

$$\alpha = P(Error\ tipo\ I) = P(Rechazar\ H_0 \mid H_0\ cierta) = P\left(T(\boldsymbol{X}) \in R_c \,\middle|\, T \sim Bin\left(n, \frac{1}{2}\right)\right)$$

Teorema del límite central para la Bin(n,1/2)

$$= P\left(|T(\boldsymbol{X}) - n/2| > a \,\middle|\, T \sim Bin\left(n, \frac{1}{2}\right)\right) = P\left(\left|\frac{T(\boldsymbol{X}) - n/2}{\sqrt{n\frac{1}{2}\left(1-\frac{1}{2}\right)}}\right| > \frac{a}{\sqrt{n\frac{1}{2}\left(1-\frac{1}{2}\right)}} \,\middle|\, T \sim Bin\left(n, \frac{1}{2}\right)\right) \approx P\left(|Z| > \frac{2a}{\sqrt{n}}\right)$$

$$\rightarrow \quad P\left(|Z| \leq \frac{2a}{\sqrt{n}}\right) \approx 1 - \alpha = 1 - 0{,}1 = 0{,}9 \quad \rightarrow \quad \frac{2a}{\sqrt{n}} \approx 1{,}645 \quad \rightarrow \quad a \approx 1{,}645\,\frac{\sqrt{100.000}}{2} \approx 260{,}097$$

Regla de decisión: Si aplicamos la metodología basada en la región crítica, es necesario evaluar la cantidad en términos de la cual hemos escrito las regions

$$|T(\boldsymbol{x}) - 100.000/2| = |50.347 - 50.000| = 347 > a \quad \rightarrow \quad T(\boldsymbol{x}) \in R_c \quad \rightarrow \textbf{Se rechaza } H_0$$

Si se aplica la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } rechazadora\ como\ \boldsymbol{x} \mid H_0\ cierta) = P\left(|T(\boldsymbol{X}) - n/2| \geq |T(\boldsymbol{x}) - n/2| \mid T(\boldsymbol{X}) \sim Bin\left(n, \frac{1}{2}\right)\right)$$

$$= P\left(\left|\frac{T(\boldsymbol{X}) - n/2}{\sqrt{n\frac{1}{2}\left(1-\frac{1}{2}\right)}}\right| \geq \left|\frac{T(\boldsymbol{x}) - n/2}{\sqrt{n\frac{1}{2}\left(1-\frac{1}{2}\right)}}\right| \,\middle|\, T(\boldsymbol{X}) \sim Bin\left(n, \frac{1}{2}\right)\right) \approx P\left(|Z| \geq \left|\frac{50347 - 50000}{\sqrt{100000\,\frac{1}{2}\left(1-\frac{1}{2}\right)}}\right|\right)$$

$$= P(|Z| \geq 2{,}19) = 2\,P(Z \leq -2{,}19) = 2 \cdot 0{,}0143 = 0{,}0248 \quad \rightarrow \quad pV < 0{,}1 = \alpha \quad \rightarrow \textbf{Se rechaza } H_0$$

Nota: (1) En este ejercicio, los tres contrastes proporcionan la misma decisión, pero en otros puede no ser así. (2) Con dos clases, el contraste ji-cuadrado no distingue dos distribuciones tales que las dos probabilidades son (½, ½), esto es, en este caso el contraste proporciona una decisión acerca de la simetría de la distribución. (3) Nótese que en este caso el contraste paramétrico y el contraste de los signos son esencialmente el mismo y se obtiene el mismo nivel crítico o p-valor.

Contrastes paramétricos

Ejercicio 1ch-p

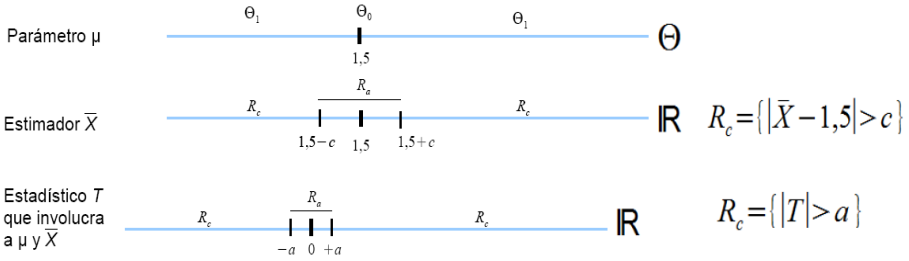
La vida media de un máquina, en años, sigue una distribución normal con varianza igual a 4. Una muestra aleatoria simple de tamaño 100 proporciona una media muestral igual a 1,3 años. Contrastar la hipótesis nula de que la media poblacional es igual a 1,5 años, aplicando un contraste bilateral de 5 por ciento de nivel de significancia.

Nivel de significancia y enunciado de las hipótesis: $\alpha = 0{,}05$ y, para un contraste bilateral,

$$H_0:\ \mu = 1{,}5 \quad \text{y} \quad H_1:\ \mu \neq 1{,}5$$

Estadístico adecuado y región crítica: Hay una población normal con varianza conocida, por lo que el estadístico incluido abajo es seleccionado. Para determinar la región crítica, bajo H_0, son necesarias su forma y los valores críticos. Es muy útil dibujar el espacio paramétrico y el espacio donde toma valores del estadístico (si se desea, también el espacio donde toma valores el estimador del parámetro):

$$T(\boldsymbol{X};\mu) = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$



Los valores críticos $-a$ y $+a$ (simétricos en este caso, dado que también lo es la distribución normal estándar) se encuentran aplicando la definición de error de tipo I:

$$\alpha = P(\textit{Error tipo I}) = P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta}) = P(T_\theta(\boldsymbol{X}) \in R_c \mid \theta \in \Theta_0) = P(|Z| > a)$$

$$\rightarrow \quad a = z_{\alpha/2} = 1{,}96 \quad \rightarrow \quad R_c = \{|T| > 1{,}96\}.$$

Regla de decisión: Para tomar la decisión final sobre las hipótesis, hay dos metodologías disponibles. Para aplicar la primera, se evalúa el estadístico T en la muestra específica $\boldsymbol{x} = (x_1,\ldots,x_{100})$:

$$T(\boldsymbol{x};\mu=1.5) = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{x}-1.5}{\sqrt{\frac{4^2}{100}}} = \frac{1.3-1.5}{\sqrt{\frac{4^2}{100}}} = \frac{-0.2\cdot 10}{4} = -\frac{1}{2} \quad \rightarrow \quad T(\boldsymbol{x};\mu=1{,}5) \notin R_c \quad \rightarrow \textbf{No se rechaza } H_0$$

Siempre se calcula la región crítica, por lo que aplicar esta metodología es fácil. La segunda metodología requiere el cálculo del nivel crítico o p-valor (p-*value*), que es por definición una probabilidad:

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta}) = P(|T(\boldsymbol{X})| \geq |T(\boldsymbol{x})| \mid \mu = 1{,}5)$$

$$= P(|Z| \geq |-0{,}5|) = 1 - P(|Z| < 0{,}5) = 1 - 0{,}3830 = 0{,}617 \quad \rightarrow \quad pV = 0{,}617 > 0{,}05 = \alpha \quad \rightarrow \textbf{No se rechaza } H_0$$

Esta última metodología proporciona la misma decisión final más información adicional sobre el soporte que la muestra $\boldsymbol{x}$ ha dado a la hipótesis nula. Como pV es bastante mayor que α, el soporte is fuerte en este caso.

Ejercicio 2ch-p

Dados 25 datos de una población normal, la información muestral se resume en

$$\sum_{i=1}^{25} x_i = 105 \qquad y \qquad \sum_{i=1}^{25} x_i^2 = 579{,}24$$

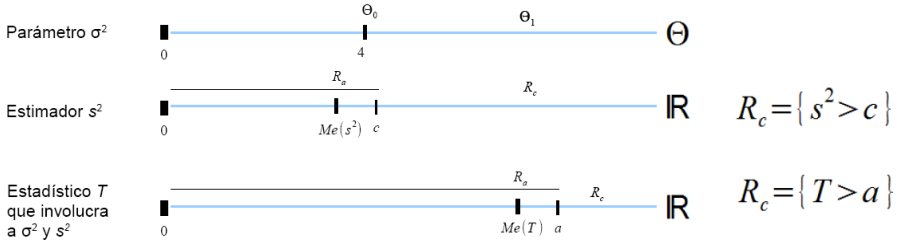
(a) ¿Debería ser rechazada la hipótesis H_0: $\sigma^2 = 4$ cuando H_1: $\sigma^2 > 4$ y $\alpha = 0{,}05$?
(b) ¿Y si H_1: $\sigma^2 \neq 4$?

(a) Hipótesis alternativa unilateral

Nivel de significancia e hipótesis: $\alpha = 0{,}05$, H_0: $\sigma^2 = 4$ y H_1: $\sigma^2 > 4$

Estadístico adecuado y región crítica: Hay una población normal con media desconocida. Para determinar la región crítica, bajo H_0, son necesarios su forma y valores críticos. Es muy útil representar el espacio paramétrico y el espacio donde toma valores el estadístico (si se desea, también el espacio donde toma valores el estimador del parámetro):

$$T(\boldsymbol{X};\sigma) = \frac{n s^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$



El valor crítico $+a$ se calcula aplicando la definición de error de tipo I:

$$\alpha = P(\text{Error tipo I}) = P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(T_0(\boldsymbol{X}) \in R_c \mid \theta \in \Theta_0) = P(\chi^2_{25-1} > a) = 1 - P(\chi^2_{25-1} \leq a)$$

$$\rightarrow \quad P(\chi^2_{24} \leq a) = 1 - \alpha = 1 - 0{,}05 = 0{,}95 \quad \rightarrow \quad a = 36{,}4 \quad \rightarrow \quad R_c = \{T > 36{,}4\}$$

Regla de decisión: Para tomar la decisión final sobre las hipótesis, hay disponibles dos metodologías. Para aplicar la primera, el estadístico T es evaluado para la muestra específica $\boldsymbol{x}$:

$$T(\boldsymbol{x}) = \frac{25\left[\frac{1}{25}\sum x_i^2 - \left(\frac{1}{25}\sum x_i\right)^2\right]}{4} = \frac{25 \cdot 5{,}53}{4} = 34{,}56 \quad \rightarrow \quad T(\boldsymbol{x}) \notin R_c \quad \rightarrow \textbf{No se rechaza } H_0$$

Para calcular la varianza, la propiedad general $s^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$ ha sido utilizada. La segunda metodología requiere el cálculo del nivel crítico o p-valor:

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta}) = P(T(\boldsymbol{X}) \geq T(\boldsymbol{x})) = P(\chi^2_{24} \geq 34{,}56)$$

$$= 1 - P(\chi^2_{24} < 34{,}56) = 0{,}075 \quad \rightarrow \quad pV = 0{,}075 > 0{,}05 = \alpha \quad \rightarrow \textbf{No se rechaza } H_0$$

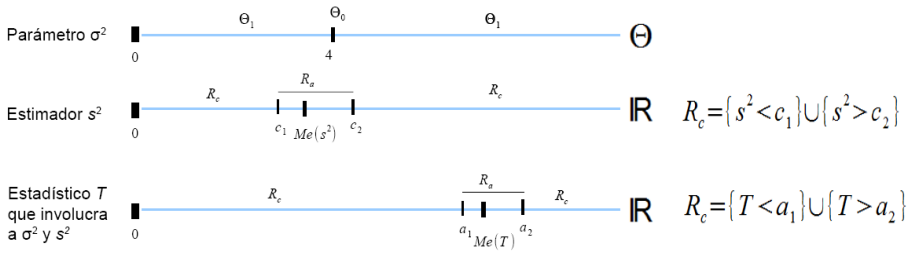
`Con el código 1-pchisq(34.56, 24)` en el lenguaje de programación *R*, dado que el valor 34,56 no está en las tablas que tenemos.

(b) Hipótesis alternativa bilateral

Nivel de significancia e hipótesis: $\alpha = 0{,}05$, H_0: $\sigma^2 = 4$ y H_1: $\sigma^2 \neq 4$

Estadístico adecuado y región crítica: Ahora

$$T(\boldsymbol{X};\sigma)=\frac{n s^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$$



Los valores críticos a_1 y a_2 se calculan aplicando la definición de error de tipo I:

$$\alpha = P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta}) = P(T_\theta(\boldsymbol{X})\in R_c \mid \theta\in\Theta_0) = P(\chi^2_{24}< a_1)+ P(\chi^2_{24}> a_2)$$

Dado que hay infinitos pares de cuantiles tales que $P(a_1<\chi^2_{24}<a_2)=1-\alpha$, se considera por convenio el que determina colas de α/2. Entonces

$$\left.\begin{array}{lll}\frac{\alpha}{2}=P(\chi^2_{24}<a_1) & \rightarrow & a_1=12{,}4\\ \frac{\alpha}{2}=P(\chi^2_{24}>a_2) & \rightarrow & a_2=39{,}4\end{array}\right\} \quad\rightarrow\quad R_c=\{T<12{,}4\}\cup\{T>39{,}4\}$$

Regla de decisión: Si se evalúa el estadístico T en la muestra particular $\boldsymbol{x}$:

$$T(\boldsymbol{x})=34{,}56 \;\rightarrow\; T(\boldsymbol{x})\notin R_c \;\rightarrow \textbf{No se rechaza } H_0$$

Para basar la decisión en el nivel crítico o p-valor, dos veces la probabilidad de la cola determinada por $T(\boldsymbol{x})$:

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ true}) = 2\cdot P(T(\boldsymbol{X})\geq T(\boldsymbol{x})) = 2\cdot P(\chi^2_{24}\geq 34{,}56)$$

$$=2[1-P(\chi^2_{24}<34{,}56)]=2\cdot 0{,}075=0{,}15 \;\rightarrow\; pV=0{,}15>0{,}05=\alpha \;\rightarrow \textbf{No se rechaza } H_0$$

Con el código `1-pchisq(34.56, 24)` en el lenguaje de programación *R*, dado que el valor 34,56 no está en las tablas que tenemos

Nota: Dados H_0 y α, se puede llegar a decisiones diferentes para los contrastes unilateral y bilateral; es por esto por lo que describir los detalles del entorno de trabajo tiene gran importancia en Estadística.

Ejercicio 3ch-p

La homocedasticidad (igualdad de varianzas) de dos poblaciones biológicas debe ser estudiada. La distribución de la variable se supone que es normal e independiente en ambas poblaciones. Después de recoger información mediante muestras de tamaños $n_X = 11$, $n_Y = 10$, respectivamente, se resume en

$$S_X^2=\frac{1}{n_X-1}\sum_{i=1}^{n_X}(x_i-\bar{x})^2=6{,}8 \qquad s_Y^2=\frac{1}{n_Y}\sum_{i=1}^{n_Y}(y_i-\bar{y})^2=7{,}1$$

Para α = 0,1, contrastar:

(a) H_0: $\sigma_X = \sigma_Y$ y H_1: $\sigma_X < \sigma_Y$
(b) H_0: $\sigma_X = \sigma_Y$ y H_1: $\sigma_X > \sigma_Y$
(c) H_0: $\sigma_X = \sigma_Y$ y H_1: $\sigma_X \neq \sigma_Y$

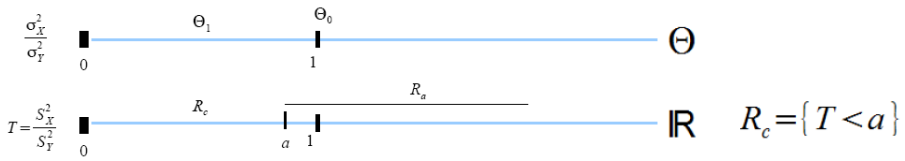
(a) Hipótesis alternativa unilateral $\sigma_X < \sigma_Y$

Nivel de significancia e hipótesis: $\alpha = 0{,}1$, H_0: $\sigma_X^2=\sigma_Y^2$ y H_1: $\sigma_X^2<\sigma_Y^2$

Estadístico adecuado y región crítica: Hay dos poblaciones normales con medias desconocidas. Para

determinar la región crítica, bajo H_0, se necesitan su forma y sus valores críticos. El estadístico de abajo se selecciona de la tabla de estadísticos (p.ej. en [2]):

$$T(\boldsymbol{X},\boldsymbol{Y})=\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}}=\frac{S_X^2}{S_Y^2}\sim F_{n_X-1,\,n_Y-1}$$



$$R_c=\{T<a\}$$

Se encuentra el valor crítico $+a$ aplicando la definición de error de tipo I:

$$\alpha=P(\textit{Error tipo I})=P(\textit{Rechazar } H_0\,|\,H_0\textit{ cierta})=P(T(\boldsymbol{X},\boldsymbol{Y})<a)=P(F_{11-1,\,10-1}<a)=P(F_{10,9}<a)$$

$$\rightarrow\quad 0.1=P(F_{10,9}<a)=P\left(F_{9,10}\geq\frac{1}{a}\right)\quad\rightarrow\quad\frac{1}{a}=2{,}35\quad\rightarrow\quad a=0{,}43\quad\rightarrow\quad R_c=\{T<0{,}43\}.$$

A partir de la definición de la distribución F de Snedecor, es fácil ver que si X sigue una $F_{n1,n2}$ entonces $1/X$ sigue una $F_{n2,n1}$. Utilizamos este truco para consultar la tabla que tenemos.

Regla de decisión: Para tomar la decisión final sobre las hipótesis, hay disponibles dos metodologías. Para aplicar la primera, se evalúa el estadístico T en las muestras específicas $\boldsymbol{x}$ y $\boldsymbol{y}$:

$$T(\boldsymbol{x},\boldsymbol{y})=\frac{S_X^2}{S_Y^2}=\frac{S_X^2}{\frac{n_Y}{n_Y-1}s_Y^2}=\frac{6{,}8}{\frac{10}{10-1}7{,}1}=0{,}86\quad\rightarrow\quad T(\boldsymbol{x})\notin R_c\quad\rightarrow\textbf{No se rechaza }H_0$$

(Para calcular la cuasivarianza S_Y^2, la propiedad general $(n-1)S^2=n\,s^2$ ha sido utilizada.) La segunda metodología requiere el cálculo del nivel crítico o p-valor:

$$pV=P(\boldsymbol{X}\text{ tan }\textit{rechazadora como }\boldsymbol{x}\,|\,H_0\textit{ cierta})=P(T(\boldsymbol{X},\boldsymbol{Y})\leq T(\boldsymbol{x},\boldsymbol{y}))=P(F_{10,9}\leq 0{,}86)=0{,}41$$

$$\rightarrow\quad pV=0{,}41>0{,}1=\alpha\quad\rightarrow\textbf{No se rechaza }H_0$$

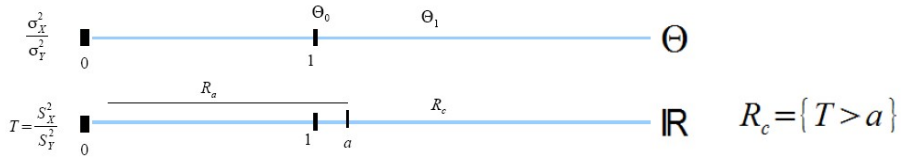
`Con el código pf(0.86, 10, 9)en el lenguaje de programación` *R*.

(b) Hipótesis alternativa unilateral $\sigma_X > \sigma_Y$

Nivel de significancia e hipótesis: $\alpha = 0{,}1$, $\quad H_0\colon \sigma_X^2=\sigma_Y^2\quad$ y $\quad H_1\colon \sigma_X^2>\sigma_Y^2$

Estadístico adecuado y región crítica:

$$T(\boldsymbol{X},\boldsymbol{Y})=\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}}=\frac{S_X^2}{S_Y^2}\sim F_{n_X-1,\,n_Y-1}$$



$$R_c=\{T>a\}$$

El valor crítico $+a$ se calcula aplicando la definición de error de tipo I:

$$\alpha=P(\textit{Error tipo I})=P(\textit{Rechazar } H_0\,|\,H_0\textit{ cierta})=P(T(\boldsymbol{X},\boldsymbol{Y})>a)=P(F_{11-1,\,10-1}>a)=P(F_{10,9}>a)$$

$$\rightarrow\quad a=2{,}42\quad\rightarrow\quad R_c=\{T>2{,}42\}.$$

Regla de decisión: Se evalúa el estadístico T en las muestras específica $\boldsymbol{x}$ y $\boldsymbol{y}$:

$$T(\boldsymbol{x},\boldsymbol{y})=0{,}86\quad\rightarrow\quad T(\boldsymbol{x})\notin R_c\quad\rightarrow\textbf{No se rechaza }H_0$$

La segunda metodología requiere el cálculo del nivel crítico o p-valor:

$$pV=P(\boldsymbol{X}\text{ tan }\textit{rechazadora como }\boldsymbol{x}\,|\,H_0\textit{ cierta})=P(T(\boldsymbol{X},\boldsymbol{Y})\geq T(\boldsymbol{x},\boldsymbol{y}))=P(F_{10,9}\geq 0{,}86)=1-0{,}41=0{,}59$$

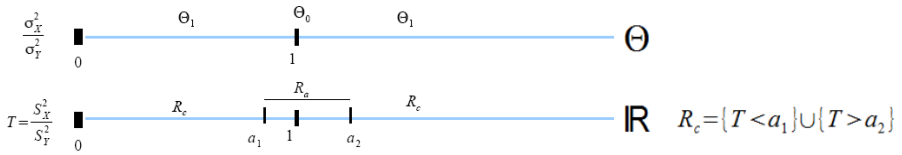
$\rightarrow \quad pV = 0{,}59 > 0{,}1 = \alpha \quad \rightarrow$ **No se rechaza** H_0

(c) Hipótesis alternativa bilateral $\sigma_X \neq \sigma_Y$

Nivel de significancia e hipótesis: $\alpha = 0{,}1$, $\quad H_0\colon \sigma_X^2 = \sigma_Y^2 \quad$ y $\quad H_1\colon \sigma_X^2 \neq \sigma_Y^2$

Estadístico apropiado y región crítica:

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X - 1,\, n_Y - 1}$$



Aplicando la definición de error de tipo I y el criterio de dejar la mitad de la probabilidad en cada cola:

$$\alpha = P(\textit{Error tipo I}) = P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta}) = P(T(\boldsymbol{X}, \boldsymbol{Y}) < a_1) + P(T(\boldsymbol{X}, \boldsymbol{Y}) > a_2)$$

$$\rightarrow \quad \left.\begin{array}{lll} \frac{\alpha}{2} = P(F_{10,9} < a_1) & \rightarrow & a_1 = 0{,}33 \\ \frac{\alpha}{2} = P(F_{10,9} > a_2) & \rightarrow & a_2 = 3{,}14 \end{array}\right\} \quad \rightarrow \quad R_c = \{T < 0{,}33\} \cup \{T > 3{,}14\}$$

donde se ha utilizado la función de cuantiles del lenguaje de programación R:

```
> qf(c(0.05 , 0.95), 10, 9)
[1] 0.3310838 3.1372801
```

Regla de decisión: El estadístico T es evaluado en las muestras concretas $\boldsymbol{x}$ y $\boldsymbol{y}$:

$$T(\boldsymbol{x}, \boldsymbol{y}) = 0{,}86 \quad \rightarrow \quad T(\boldsymbol{x}) \notin R_c \quad \rightarrow \textbf{No se rechaza } H_0$$

Para aplicar la metodología basada en el nivel crítico, calculamos `qf(0.5, 10, 9)=1.007739`, que es la mediana; como $T(\boldsymbol{x},\boldsymbol{y})$ está en la cola izquierda:

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ true}) = 2\,P(T(\boldsymbol{X}, \boldsymbol{Y}) \leq T(\boldsymbol{x}, \boldsymbol{y})) = 2 \cdot 0{,}41 = 0{,}82$$

$$\rightarrow \quad pV = 0{,}82 > 0{,}1 = \alpha \quad \rightarrow \textbf{No se rechaza } H_0$$

Nota: Por definición, el nivel crítico o p-valor toma un valor entre 0 y 1 (es una probabilidad). Si no sabemos en qué cola está $T(\boldsymbol{x},\boldsymbol{y})$ y consideramos la incorrecta, nos daremos cuenta porque el doble de la probabilidad de la cola sería mayor a 1.

Contrastes no paramétricos

Ejercicio 1ch-np

Tres productos financieros han sido comercializados y la presencia de interés *ha sido registrada para algunas personas. Es posible imaginar diferentes situaciones en las que los siguientes datos podrían ser obtenidos.*

	Producto 1	***Producto 2***	***Producto 3***	
Grupo 1	*10*	*18*	*9*	*37*
Grupo 2	*20*	*13*	*15*	*48*
	30	*31*	*24*	*85*

(a) Si 48 personas del grupo 2 fueron situados considerando la variable producto*, contrasta si esta variable sigue la distribución determinada por la muestra del grupo 1 cuando α = 0.01.*

(b) Si se entrevista a 37 personas del primer grupo y 48 personas del segundo, contrasta la homogeneidad de la distribución de la variable producto *en ambos grupos cuando α = 0,01.*

(c) La gente con interés en alguno de los productos fue clasificada después de considerar las dos variables grupo *y* producto*. Contrasta la independencia de las dos variables cuando α = 0,01.*

(a) Hipótesis y nivel de significancia: El nivel α = 0,01 está dado. Para un contraste no paramétrico de bondad de ajuste, la hipótesis nula supone que las probabilidades teóricas del segundo grupo siguen las probabilidades de la muestra del primer grupo, la referencia. Si P_i representa la variable *producto* en la población *i*-ésima,

$$H_0:\ P_2 \sim P_1 \quad \text{y} \quad H_1:\ P_2 \sim F \neq P_1$$

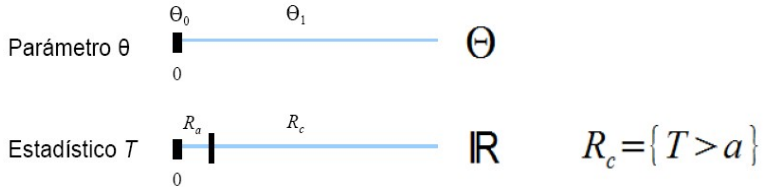
Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]),

$$T(\boldsymbol{X}) = \sum_{i=1}^{K} \frac{(N_i - \hat{e}_i)^2}{\hat{e}_i} \xrightarrow{d} \chi^2_{K-s-1}$$

con

K = 3 clases y 1 muestra

No ha debido ser estimada ninguna probabilidad, así $s = 0$



Para calcular el cuantil *a*, se aplica la definición de error de tipo I:

$$\alpha = P(\textit{Error tipo I}) = P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta}) = P(T(\boldsymbol{X}) \in R_c \mid T \approx \chi^2_2) \approx P(\chi^2_2 > a) = 1 - P(\chi^2_2 \leq a)$$

$$\rightarrow \quad P(\chi^2_2 \leq a) \approx 1 - \alpha = 1 - 0{,}01 = 0{,}99 \quad \rightarrow \quad a \approx 9{,}21$$

Regla de decisión:

La variable P_1 sigue la distribución con valores {1, 2, 3} y probabilidades {10/37, 18/37, 9/37}.

Frecuencias esperadas				
	Producto 1	Producto 2	Producto 3	
Grupo 2	e_1	e_2	e_3	48

$e_1 = 48\frac{10}{37}$ $e_2 = 48\frac{18}{37}$ $e_3 = 48\frac{9}{37}$

Frecuencias empíricas				
	Producto 1	Producto 2	Producto 3	
Grupo 2	20	13	15	48

Si aplicamos la metodología basada en la región crítica,

$$T(\boldsymbol{x}) = \frac{\left(20 - 48\frac{10}{37}\right)^2}{48\frac{10}{37}} + \frac{\left(13 - 48\frac{18}{37}\right)^2}{48\frac{18}{37}} + \frac{\left(15 - 48\frac{9}{37}\right)^2}{48\frac{9}{37}} = 9{,}34 \quad \rightarrow \quad T(\boldsymbol{x}) \in R_c \quad \rightarrow \textbf{Se rechaza } H_0$$

Si aplicamos la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta}) = P(T(\boldsymbol{X}) \geq T(\boldsymbol{x}) \mid T \approx \chi^2_2) \approx P(\chi^2_2 \geq 9{,}34)$$

$$< P(\chi^2_2 \geq 9{,}21) = 1 - P(\chi^2_2 < 9{,}21) = 1 - 0{,}99 = 0{,}01 \quad \rightarrow \quad pV < 0{,}01 = \alpha \quad \rightarrow \textbf{Se rechaza } H_0$$

9,34 no está en las tablas que tenemos, mintras que 9,21 sí está

(b) Hipótesis y nivel de significancia: El nivel α = 0,01 está dado. Para un contraste no paramétrico de homogeneidad, la hipótesis nula supone que las probabilidades de cualquier columna son las mismas para los dos grupos, esto es, son independientes del grupo o estrato. Si *G* representa la variable grupo,

$$H_0\colon\ F(x|G)=F(x) \quad \text{y} \quad H_1\colon\ F(x|G)\neq F(x)$$

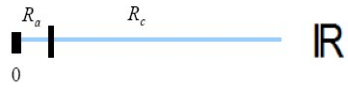
Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]),

$$T(\boldsymbol{X})=\sum_{i=1}^{L}\sum_{j=1}^{K}\frac{(N_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} \xrightarrow{d} \chi^2_{(L-1)(K-1)}$$

con

$L = 2$ grupos y $K = 3$ clases

Dos de las tres probabilidades deben ser estimadas, así $s = 2$



$$R_c=\{T>a\}$$

Para calcular el cuantil a, se aplica la definición de error de tipo I:

$$\alpha=P(\textit{Error tipo I})=P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta})=P(T(\boldsymbol{X})\in R_c \mid T\approx\chi^2_2)\approx P(\chi^2_2>a)=1-P(\chi^2_2\leq a)$$

$$\rightarrow \quad P(\chi^2_2\leq a)\approx 1-\alpha=1-0{,}01=0{,}99 \quad \rightarrow \quad a\approx 9{,}21$$

Regla de decisión: Se supone una distribución subyacente, aunque no específica, por lo que las probabilidades son estimadas directamente a partir de la información muestral.

Frecuencias esperadas

	Producto 1	Producto 2	Producto 3	
Grupo 1	$\hat{e}_{11}$	$\hat{e}_{12}$	$\hat{e}_{13}$	37
Grupo 2	$\hat{e}_{21}$	$\hat{e}_{22}$	$\hat{e}_{23}$	48
				85

$\hat{e}_{11}=37\frac{30}{85}$ $\quad \hat{e}_{12}=37\frac{31}{85}$ $\quad \hat{e}_{13}=37\frac{24}{85}$

$\hat{e}_{21}=48\frac{30}{85}$ $\quad \hat{e}_{22}=48\frac{31}{85}$ $\quad \hat{e}_{23}=48\frac{24}{85}$

Frecuencias empíricas

	Producto 1	Producto 2	Producto 3	
Grupo 1	10	18	9	37
Grupo 2	20	13	15	48
	30	31	24	85

$\hat{p}_{\cdot 1}=\frac{30}{85}$ $\quad \hat{p}_{\cdot 2}=\frac{31}{85}$ $\quad \hat{p}_{\cdot 3}=\frac{24}{85}$

Si aplicamos la metodología basada en la región crítica,

$$T(\boldsymbol{x})=\frac{\left(10-37\frac{30}{85}\right)^2}{37\frac{30}{85}}+\cdots+\frac{\left(15-48\frac{24}{85}\right)^2}{48\frac{24}{85}}=4{,}29 \quad \rightarrow \quad T(\boldsymbol{x})\notin R_c \quad \rightarrow \textbf{No se rechaza } H_0$$

Si aplicamos la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta})=P(T(\boldsymbol{X})\geq T(\boldsymbol{x}) \mid T\approx\chi^2_2)\approx P(\chi^2_2\geq 4{,}29)$$

$$=1-P(\chi^2_2<4{,}29)>1-P(\chi^2_1<4{,}61)=1-0{,}9=0{,}1 \quad \rightarrow \quad pV>0{,}1>0{,}01=\alpha \quad \rightarrow \textbf{No se rechaza } H_0$$

4,29 no está en la tabla que tenemos, mientras que 4,61 sí está

(c) Hipótesis y nivel de significancia: El nivel $\alpha = 0{,}01$ está dado. Para un contraste no paramétrico de independencia, la hipótesis nula supone que las probabilidades de cada celda es el producto de las probabilidades de su fila y columna,

$$H_0\colon\ f(x,y)=f(x)f(y) \quad \text{y} \quad H_1\colon\ f(x,y)\neq f(x)f(y)$$

Estadístico y región crítica: A partir de la tabla de estadísticos (p.ej. en [2]),

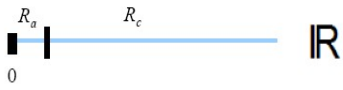
$$T(\boldsymbol{X})=\sum_{i=1}^{L}\sum_{j=1}^{K}\frac{(N_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} \xrightarrow{d} \chi^2_{(L-1)(K-1)}$$

Parámetro θ 

con

$L = 2$ y $K = 3$ clases

Una y dos probabilidades deben ser estimadas, así $s = 3$

Estadístico T 

$$R_c=\{T>a\}$$

Para calcular el cuantil a, se aplica la definición de error de tipo I:

$$\alpha=P(\textit{Error tipo I})=P(\textit{Rechazar } H_0 \mid H_0 \textit{ cierta})=P(T(\boldsymbol{X})\in R_c \mid T\approx\chi^2_2)\approx P(\chi^2_2>a)=1-P(\chi^2_2\le a)$$

$$\rightarrow \quad P(\chi^2_2\le a)\approx 1-\alpha=1-0{,}01=0{,}99 \quad \rightarrow \quad a\approx 9{,}21$$

Regla de decisión: Se supone una distribución subyacente, aunque no específica, por lo que las probabilidades son estimadas directamente a partir de la información muestral.

Frecuencias esperadas

	Producto 1	Producto 2	Producto 3	
Grupo 1	$\hat{e}_{11}$	$\hat{e}_{12}$	$\hat{e}_{13}$	37
Grupo 2	$\hat{e}_{21}$	$\hat{e}_{22}$	$\hat{e}_{23}$	48
				85

$\hat{e}_{11}=85\frac{37}{85}\frac{30}{85}$ $\hat{e}_{12}=85\frac{37}{85}\frac{31}{85}$ $\hat{e}_{13}=85\frac{37}{85}\frac{24}{85}$

$\hat{e}_{21}=85\frac{48}{85}\frac{30}{85}$ $\hat{e}_{22}=85\frac{48}{85}\frac{31}{85}$ $\hat{e}_{23}=85\frac{48}{85}\frac{24}{85}$

Frecuencias empíricas

	Producto 1	Producto 2	Producto 3		
Grupo 1	10	18	9	37	$\hat{p}_{1-}=\frac{37}{85}$
Grupo 2	20	13	15	48	$\hat{p}_{2-}=\frac{48}{85}$
	30	31	24	85	

$\hat{p}_{-1}=\frac{30}{85}$ $\hat{p}_{-2}=\frac{31}{85}$ $\hat{p}_{-3}=\frac{24}{85}$

Si aplicamos la metodología basada en la región crítica,

$$T(\boldsymbol{x})=\frac{\left(10-37\frac{30}{85}\right)^2}{37\frac{30}{85}}+\cdots+\frac{\left(15-48\frac{24}{85}\right)^2}{48\frac{24}{85}}=4{,}29 \quad \rightarrow \quad T(\boldsymbol{x})\notin R_c \quad \rightarrow \textbf{No se rechaza } H_0$$

Si aplicamos la metodología basada en el nivel crítico o p-valor,

$$pV = P(\boldsymbol{X} \text{ tan } \textit{rechazadora como } \boldsymbol{x} \mid H_0 \textit{ cierta})=P(T(\boldsymbol{X})\ge T(\boldsymbol{x}) \mid T\approx\chi^2_2)\approx P(\chi^2_2\ge 4{,}29)$$

$$=1-P(\chi^2_2<4{,}29)>1-P(\chi^2_1<4{,}61)=1-0{,}9=0{,}1 \quad \rightarrow \quad pV>0{,}1>0{,}01=\alpha \quad \rightarrow \textbf{No se rechaza } H_0$$

Nota: El contraste de homogeneidad puede ser visto como un caso particular del contraste de independencia donde una variable, digamos G, indica la pertenencia al grupo o estrato y, al mismo tiempo, el número de elementos en cada muestra ha sido fijado, lo que puede verse como restricciones donde se condiciona la distribución conjunta a ciertos valores para G; esto implica que las probabilidades «se estiman automáticamente». Nótese que los resultados numéricos y la decisión son los mismos en ambos tipos de contraste. Por otro lado, el contraste de bondad de ajuste puede ser visto como un caso particular del contraste de homogeneidad con dos muestras donde una de ellas determina el modelo de referencia para la hipótesis nula (un vector de frecuencias determina un vector de probabilidades).

Nota: En este ejercicio, la independencia y homogeneidad no han sido rechazadas, mientras que la hipótesis que supone que la variable producto sigue en la población 2 la distribución determinada por la muestra del grupo 1. Por otro lado, la distribución determinada por una muestra, involucrada en (a), es en general diferente de la distribución subyacente común supuesta, involucrada en (b) y (c), que es estimada usando las muestras de ambos grupos. Entonces, esta distribución subyacente «está entre las dos muestras», lo que puede justificar las decisiones en (a), (b) y (c); de hecho, en este caso concreto el grupo 2 tiene mayor peso al determinar esa distribución por tener más elementos.

Nota: En la práctica, para las estimaciones $\hat{e}_{ij}$ puede aplicarse la misma regla mnemotécnica tanto en el contraste de homogeneidad como en el de independencia: para cada posición, multiplicar por las frecuencias absolutas de la fila y la columna y dividir por el número total de elementos n.

Referencias

[1] Casado, D. *Herramientas básicas de Matemáticas*, http://www.casado-d.org/edu/HerramientasBasicasDeMatematicas.pdf

[2] Casado, D. *Inferencia estadística*,
http://www.casado-d.org/edu/docencia.html#MatematicasIntermedias-Estadistica-InferenciaEstadistica

Universidad Complutense de Madrid
∟ Facultad de Ciencias Económicas y Empresariales
∟ Departamento de Estadística e Investigación Operativa II
∟ David Casado de Lucas

29 de octubre del 2012